ΕΘΝΙΚΌ ΜΕΤΣΌΒΙΟ ΠΟΛΥΤΕΧΝΕΊΟ
ΣΧΟΛΉ ΗΛΕΚΤΡΟΛΌΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σχεδίαση και Υλοποίηση Συστήματος Σύστασης Βασισμένου σε Γνώση για την Αντιμετώπιση του Προβλήματος Ψυχρής Εκκίνησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Στέφανος Παναγιώτης Ν Αργυρίου

Επιβλέπων : **Γεώργιος Στάμου**
Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2018

ΕΘΝΙΚΌ ΜΕΤΣΌΒΙΟ ΠΟΛΥΤΕΧΝΕΊΟ
ΣΧΟΛΉ ΗΛΕΚΤΡΟΛΌΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σχεδίαση και Υλοποίηση Συστήματος Σύστασης Βασισμένου σε Γνώση για την Αντιμετώπιση του Προβλήματος Ψυχρής Εκκίνησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Στέφανος Παναγιώτης Ν Αργυρίου

Επιβλέπων : **Γεώργιος Στάμου**
Αναπληρωτής Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Μαρτίου 2018.

Γεώργιος Στάμου Αναπληρωτής Καθηγητής ΕΜΠ	Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής ΕΜΠ	Κωνσταντίνα Νικήτα Καθηγήτρια ΕΜΠ

Αθήνα, Μάρτιος 2018

Στέφανος Παναγιώτης Ν Αργυρίου

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

Περίληψη

Η συνεχής εξέλιξη του διαδικτύου και η διάδοση των μέσων μαζικής δικτύωσης και επικοινωνίας έχουν καταστήσει ευκολότερη από ποτέ τη δήλωση και συλλογή πληροφοριών σχετικά με τα ενδιαφέροντα, τις ασχολίες και τις καταναλωτικές συνήθειες κάθε χρήστη ξεχωριστά. Σε αυτά τα πλαίσια, ο τομέας των συστήματων σύστασης εξελίχθηκε και συνεχίζει να εξελίσσεται με αλματώδεις ρυθμούς ακόμα και εν έτει 2017, 10 χρόνια μετά τη διεξαγωγή του Netflix Prize Event.

Σκοπός της παρούσης εργασίας αποτελεί η ανάλυση, σχεδίαση και δημιουργία ενός συστήματος σύστασης βασισμένου σε γνώση για την αντιμετώπιση του προβλήματος της ψυχρής εκκίνησης, κατάσταση η οποία μειώνει την αποτελεσματικότητα ακόμα και των πιο πετυχημένων διαδικτυακών συστημάτων σύστασης.

Αρχικά, γίνεται μία εισαγωγή στον τομέα των συστημάτων σύστασης και στον τρόπο με τον οποίο πραγματοποιείται η οντολογική αναπαράσταση γνώσης και οι αντίστοιχες συλλογιστικές διαδικασίες. Ιδιαίτερη έμφαση δίνεται στον τρόπο με τον οποίο τα παραπάνω πεδία εκφράζονται στον Παγκόσμιο Ιστό. Με βάση τα βασικά χαρακτηριστικά και τη θεωρητική θεμελίωση των προαναφερθέντων τομέων επιχειρήθηκε η δημιουργία ενός όσο το δυνατόν πιο αποτελεσματικού συστήματος σύστασης βασισμένου σε γνώση.

Σε πρώτο στάδιο σχεδιάστηκε η οντολογία που τροφοδοτείται στο σύστημα για την εξαγωγή των συστάσεων. Πρόκειται για μία κινηματογραφική οντολογία που αντλήθηκε από το διαδίκτυο και πιο συγκεκριμένα από τις βάσεις δεδομένων του Movielens και του IMDB.

Κατόπιν, αναλύθηκε η λειτουργία της μηχανής συστάσεων για την παροχή προτάσεων στις περιπτώσεις που ο χρήστης εισάγει ως ενδιαφέροντά του μία ή περισσότερες ταινίες. Έγινε ακόμη, εκτενής προσπάθεια για τη δημιουργία ενός περιβάλλοντος διεπαφής με το χρήστη που να επιτελεί τους στόχους του συστήματος χωρίς να εμβαθύνει περαιτέρω σε λειτουργίες που δε συνδέονταν άμεσα με το σύστημα σύστασης.

Περίληψη

Στο τελευταίο κεφάλαιο γίνεται σύγκριση των αποτελεσμάτων του συστήματος που κατασκευάστηκε με τους αλγορίθμους για εξαγωγή προτάσεων που χρησιμοποιούνται εκτενώς από το Apache Mahout.

Λέξεις-Κλειδιά: Συστήματα Σύστασης Βασισμένα σε Γνώση, Συνεργατικό Φιλτράρισμα, Συστήματα Σύστασης Βασισμένα στο Περιεχόμενο, Επεξεργασία Γνώσης, Οντολογία, Αντικείμενο, Άτομο, Έννοια, Κλάση, Ρόλος, Συζευκτικό Ερώτημα, Υπαρξιακοί Δείκτες, Κριτική Αποτελεσμάτων, Terminological Box, Assertion Box, Συλλογιστική, Ενιαίος Προσδιοριστής Πόρου, Ενιαίος Εντοπιστής Πόρου, Οντολογική Γλώσσα Διαδικτύου, Αντικειμενοστραφής Σχεσιακή Χαρτογράφηση, Γράφος, Κόμβος, Ομοιότητα, Ομοιότητα Jaccard, Πρόβλημα Κάλυψης Συνόλου, Σελίδες JSP, Apache Mahout

Summary

The constant evolution of the Web and social media have made the collection of data, hobbies and consumer habits easier than ever for each unique user. Under this spectrum, the field of recommender systems has evolved and keeps evolving even in 2017, 10 years after the Netflix Prize Event.

The goal of this project is the analysis, the design and the creation of a Knowledge Based Recommender System to deal with the Cold Start Problem, a condition that lowers the effectiveness of even the most successful Recommender Systems.

In the beginning, a short introduction on the field of Recommender Systems and the way the ontologic representation of kowledge and the corresponding reasoning techniques. Special emphasis is given to the way knowledge is represented on the Web. Based on the study of the aforementioned fields and the theoritical foundations of the field of Knowledge Engineering a Knowledge Based Recommender system was created.

At first, the associated ontology that supplied the system with information was created. It was a cinematographic ontology that derived from the Web and more precisely from the databases of Movielens and iMDB.

Then, the responsiveness of the recommender engine was analysed when the user would express interest in one or multiple movies. There was also significant effort in the creation of a user interface that would allow the user to interact with the recommendations, without however offering more functionality than just the presentaton and selection of movies.

In the last chapter, the results of already existing recommender systems were compared with the results of the newly created Knowledge Based Recommender System.

Keywords: Knowledge Based Recommender System, Collaborative Filtering, Content-based Recommender System, Knowledge Engineering, Ontology, Item, Individual, Concept, Class, Role, Conjumctive Query, Existential Operator, Result Critique, Terminological Box, Assertion Box, Reasoning, Uniform Resource Identifier, Uniform Resource Locator, Web Ontology Language, Object Relational Mapping, Graph, Node, Similarity, Jaccard Similarity, Cover Set Problem, JSP Pages, Apache Mahout

Ευχαριστίες

Ειδικές ευχαριστίες για την ολοκλήρωση της εργασίας και την υλοποίηση της αντίστοιχης εφαρμογής πρέπει να απονεμηθούν στον καθηγητή Στάμου Γιώργο, υπεύθυνο της πτυχιακής εργασίας, όπως και σε όλους τους υπόλοιπους καθηγητές του τομέα Ευφυών Συστημάτων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου.

Επίσης, τίποτα δε θα ήταν δυνατό χωρίς τη συμβολή του Διδακτορικού φοιτητή Γιαζίτζογλου Μιχάλη, επιβλέποντα της παρούσης πτυχιακής εργασίας, ο οποίος συνεισέφερε με παροχή της δικιάς του επιστημονικής έρευνας και επιτρέποντας απρόσκοπτη χρήση των αποτελεσμάτων που έχει αποκομίσει από τη μέχρι τώρα Διδακτορική του πορεία.

Το λογότυπο της εφαρμογής και το γενικότερο ύφος της ιστοσελίδας εμπνεύστηκε και σχεδίασε ο μαθητευόμενος graphic designer Κολόκας Γεώργιος.

Ευχαριστίες πρέπει να απονεμηθούν τέλος στους Κονιδάρη Φίλιππο, Αλεξανδρίδη Γεώργιο και Κάτσιο Κωνσταντίνο για τις δημιουργικές τους προτάσεις που ώθησαν προς βελτίωση το αποτέλεσμα της εργασίας.

Περιεχόμενα

Περιεχόμενα

Περιεχόμενα

Η σελίδα αυτή είναι σκόπιμα λευκή.

1 Εισαγωγή στα Συστήματα Σύστασης

The recommendation I always give people when I 'm mentoring them is, do not be afraid to take risks.

- Ginni Rometty, CEO of IBM

Περίληψη. Σε αυτό το κεφάλαιο παρουσιάζονται τα βασικά στοιχεία που συνιστούν τον τομέα των συστημάτων σύστασης. Αρχικά περιγράφεται η έννοια του συστήματος σύστασης και ορίζεται το πρόβλημα της οργανωμένης παροχής συστάσεων, ενώ εν συνεχεία γίνεται αναλυτική αναφορά στους στόχους των συστημάτων αυτού του είδους. Το εναπομείναν μέρος του κεφαλαίου κατηγοριοποιεί τα συστήματα σύστασης ανάλογα με τη λογική που ακολουθεί η διαδικασία παραγωγής προτάσεων και παρέχει ορισμένες θεμελιώδεις πληροφορίες για τον τρόπο υλοποίησης κάθε μίας από αυτές.

1.1 Βασικές Αρχές των Συστημάτων Σύστασης

Η συνεχώς αυξανόμενη σημασία του Διαδικτύου δημιούργησε ένα μέσο για ηλεκτρονικές και επιχειρησιακές συναλλαγές που αποτελέσαν καταλύτες στην ανάπτυξη της τεχνολογίας των Συστημάτων Σύστασης[1]. Βασικός παράγοντας υπό αυτό το πρίσμα είναι η ευκολία με την οποία το Διαδίκτυο επιτρέπει στους χρήστες να μεταδώσουν πληροφορίες για τα αντικείμενα και τις ασχολίες που τους ενδιαφέρουν ή δεν τους ενδιαφέρουν. Για παράδειγμα, έστω το σενάριο μιας υπηρεσίας παροχής περιεχομένου, όπως το Netflix. Σε αυτές τις περιπτώσεις, οι χρήστες είναι πολύ εύκολο να παρέχουν πληροφορίες για τα ενδιαφέροντά τους μέσω ενός κλικ του ποντικιού τους. Μία τυπική μεθοδολογία για την προσφορά τέτοιου είδους πληροφορίας αποτελεί η παροχή κάποιου rating, με το οποίο οι χρήστες δίνουν αριθμητικές τιμές μέσω ενός συγκεκριμένου συστήματος αξιολόγησης (για παράδειγμα κλίμακα 1-5 αστέρων) για να δηλώσουν κατά πόσο τους ενδιαφέρουν συγκεκριμένα αντικείμενα.

Άλλες μορφές παροχής πληροφοριών ενδιαφέροντος μπορεί να μην είναι τόσο άμεσες, αλλά είναι πολύ πιο εύκολα συλλέξιμες σε ένα δικτυο-κεντρικό περιβάλλον. Σε αυτά τα πλαίσια, η απλή πράξη της αγοράς ενός αντικειμένου – για παράδειγμα από το Amazon – δηλώνει ξεκάθαρα το ενδιαφέρον του χρήστη για αντικείμενα της συγκεκριμένης κατηγορίας. Η βασική ιδέα των

συστημάτων σύστασης είναι η χρησιμοποίηση όλων αυτών των δεδομένων προκειμένου να εντοπίσουν και να εξάγουν τα ενδιαφέροντα του χρήστη. Η οντότητα στην οποία οι συστάσεις απευθόνονται αναφέρεται ως “χρήστης” και οι ταινίες, τα προϊόντα κλπ τα οποία δίνονται ως συστάσεις στο χρήστη ονομάζονται “αντικείμενα”. Επομένως, η ανάλυση των συστάσεων συνήθως βασίζεται σε προηγούμενες αλληλεπιδράσεις μεταξύ του χρήστη και των αντικειμένων ενδιαφέροντος, καθώς προηγούμενα ενδιαφέροντα και κλίσεις αποτελούν καλούς δείκτες μελλοντικών επιλογών. Μία σημαντική εξαίρεση στη συγκεκριμένη πρακτική αποτελούν τα Συστήματα Σύστασης βασισμένα σε Γνώση, όπου οι συστάσεις γίνονται με βάση τον προσδιορισμό συγκεκριμένων *απαιτήσεων* που ορίζονται από το χρήστη και όχι με βάση προηγούμενες αλληλεπιδράσεις του χρήστη με το σύστημα.

Ποια είναι λοιπόν η βασική αρχή στην οποία στηρίζεται κάθε αλγόριθμος σύστασης; Πρόκειται για την παραδοχή ότι υπάρχουν σημαντικές εξαρτήσεις ανάμεσα στις δραστηριότητες που αφορούν συγκεκριμένους χρήστες και συγκεκριμένης κατηγορίας αντικείμενα. Για παράδειγμα, ένας χρήστης που ενδιαφέρεται για Ιστορικά Ντοκυμαντέρ έχει μεγαλύτερες πιθανότητες να ενδιαφερθεί στη συνέχεια και για άλλα Ιστορικά Ντοκυμαντέρ ή Επιμορφωτικά Προγράμματα, από ότι ας πούμε για μία ταινία Δράσης. Σε πολλές περιπτώσεις, διάφορες κατηγορίες αντικειμένων μπορεί να εμφανίσουν σημαντικές συσχετίσεις, γεγονός το οποίο οδηγεί σε πιο ακριβείς συστάσεις. Εναλλακτικά, συσχετίσεις μεταξύ αντικειμένων μπορεί να εντοπίζονται και στις πιο μικρές λεπτομέρειες που τα χαρακτηρίζουν και να μην έχουν καμία σχέση με τις ευρύτερες κατηγορίες στις οποίες ανήκουν. Τέτοιου είδους εξαρτήσεις μπορούν να γίνουν αντιληπτές μέσω μίας δεδομενο-κεντρικής ανάλυσης μέσω του πίνακα βαθμολογιών, ο οποίος χρησιμοποιείται για την πρόβλεψη των προτιμήσεων συγκεκριμένων χρηστών. Όσο μεγαλύτερος είναι ο αριθμός των βαθμολογημένων αντικειμένων που είναι διαθέσιμα για το συγκεκριμένο χρήστη, τόσο καλύτερη είναι η δυνατότητα πρόβλεψης των μελλοντικών ενδιαφερόντων του χρήστη και κατ' επέκταση τόσο καλύτερη η ποιότητα των προτεινόμενων συστάσεων. Πολλά διαφορετικά μοντέλα έχουν εφαρμοστεί για την επίτευξη αυτού του στόχου. Για παράδειγμα, η συλλογική αγοραστική ή βαθμολογική συμπεριφορά των χρηστών μπορεί να επεξεργασθεί με σκοπό τη δημιουργία ομάδων ατόμων με παρόμοια ενδιαφέροντα. Μετά από αλληλεπίδραση του χρήστη με το σύστημα, αυτός

τοποθετείται σε μία από αυτές τις ομάδες και οι συστάσεις που πραγματοποιούνται γίνονται με βάση τα χαρακτηριστικά της ομάδας.

Η παραπάνω περιγραφή είναι βασισμένη σε μία πολύ απλή οικογένεια αλγορίθμων σύστασης, στους οποίους γίνεται αναφορά ως “μοντέλα γειτνίασης”. Αυτή η οικογένεια ανήκει σε μία ακόμα μεγαλύτερη κατηγορία μοντέλων που αποκαλούνται “μοντέλα συνεργατικού φιλτραρίσματος”. Ο όρος “συνεργατικό φιλτράρισμα” αναφέρεται στη χρήση των βαθμολογιών από πολλούς χρήστες με συνεργατικό τρόπο ώστε να προβλεφθούν βαθμολογίες που δεν βρίσκονται στο σύστημα. Στην πραγματικότητα, τα συστήματα σύστασης μπορεί να είναι πολύ πιο πολύπλοκα και να αξιοποιούν ένα μεγάλο αριθμό βοηθητικών δεδομένων εκτός από τις βαθμολογίες. Για παράδειγμα σε ένα σύσταση σύστασης βασισμένο σε περιεχόμενα, το περιεχόμενο κάθε αντικειμένου διαδραματίζει πρωταρχικό ρόλο στη διαδικασία σύστασης, στην οποία οι βαθμολογίες των χρηστών σε συνεργασία με τα χαρακτηριστικά των αντικειμένων εξετάζονται προκειμένου να παραχθούν οι προβλέψεις. Η βασική ιδέα είναι ότι τα ενδιαφέροντα του χρήστη μπορούν να μοντελοποιηθούν στη βάση των χαρακτηριστικών των αντικειμένων τα οποία ο χρήστης στο παρελθόν έχει βαθμολογήσει ή έχει αποκτήσει πρόσβαση. Μία διαφορετική δομή εμφανίζεται στα συστήματα σύστασης βασισμένα σε γνώση, στα οποία οι χρήστες διαδραστικά προσδιορίζουν τα ενδιαφέροντά τους και αυτός ο προσδιορισμός συνδυάζεται με την αποθηκευμένη γνώση του πεδίου ενδιαφέροντος για την παροχή συστάσεων. Στα πιο σύνθετα συστήματα, συμφραζόμενες και συναφείς πληροφορίες όπως η τοποθεσία του χρήστη, κοινωνικές πληροφορίες ή πληροφορίες από το Διαδίκτυο συνδυάζονται για την παραγωγή συστάσεων.

1.2 Στόχοι των Συστημάτων Σύστασης

Η αύξηση του αριθμού των πωλήσεων είναι ο βασικός στόχος ενός συστήματος σύστασης. Τα συστήματα σύστασης χρησιμοποιούνται εξάλλου κατά κόρον από τους πωλητές για τη μεγιστοποίηση των κερδών τους. Μέσω της προσεκτικής σύστασης αντικειμένων στους χρήστες, τα συστήματα σύστασης τραβούν την προσοχή των χρηστών σε συγκεκριμένα αντικείμενα. Κάτι τέτοιο, αυξάνει τις πωλήσεις και τα κέρδη για τον πωλητή. Παρόλο που ο πρωταρχικός στόχος ενός συστήματος σύστασης είναι η αύξηση των εσόδων

ενός πωλητή, ο τρόπος με τον οποίο αυτό επιτυγχάνεται συνήθως δεν είναι ορατός με την πρώτη ματιά. Προκειμένου να επιτευχθεί ο ευρύτερος επιχειρησιακός και οικονομικός στόχος της αύξησης των εσόδων οι κοινοί λειτουργικοί και τεχνικοί στόχοι των συστημάτων σύστασης είναι οι εξής:

1. Σχετικότητα: Ο πιο εμφανής λειτουργικός στόχος ενός συστήματος σύστασης είναι η σύσταση αντικειμένων που σχετίζονται με τον εκάστοτε χρήστη. Οι χρήστες έχουν μεγαλύτερη πιθανότητα να αγοράσουν ή να ενδιαφερθούν για ένα αντικείμενο το οποίο θεωρούν ενδιαφέρον. Παρόλο που η σχετικότητα είναι ο πρωταρχικός στόχος ενός συστήματος σύστασης, δεν είναι αρκετός από μόνος του.

2. Καινοτομία: Τα συστήματα σύστασης είναι πραγματικά χρήσιμα σε κάποιον χρήστη όταν προτείνουν αντικείμενα τα οποία ο χρήστης δεν έχει δει στο παρελθόν. Για παράδειγμα, δημοφιλείς ταινίες μιας κατηγορίας που ενδιαφέρει το χρήστη ελάχιστες φορές θα είναι άγνωστες σε αυτόν. Επαναλαμβανόμενες συστάσεις δημοφιλών αντικειμένων μπορεί να οδηγήσει ακόμα και σε μείωση των συνολικών πωλήσεων για τον πωλητή που αξιοποιεί το σύστημα σύστασης.

3. Εναλλακτικότητα: Μία ιδέα αρκετά σχετική με αυτή της καινοτομίας, σύμφωνα με την οποία τα προτεινόμενα αντικείμενα είναι απροσδόκητα σε σχέση με τα συνήθη ενδιαφέροντα του χρήστη. Η διαφορά με την καινοτομία είναι ότι τα αντικείμενα που προκύπτουν ως συστάσεις με βάση τις αρχές της εναλλακτικότητας είναι πραγματικά μη αναμενόμενα από το χρήστη και δεν αποτελούν απλώς αντικείμενα που ο χρήστης έχει τύχει να μην έχει ακούσει ως τότε. Για παράδειγμα, εάν ένα νέο Ινδικό εστιατόριο ανοίξει σε μία περιοχή, τότε σύμφωνα η σύσταση αυτού του εστιατορίου θα ήταν καινοτόμα για ένα χρήστη που του αρέσει το Ινδικό φαγητό, αλλά δε θα ήταν εναλλακτική. Αντίθετα, η σύσταση στον ίδιο χρήστη ενός εστιατορίου με φαγητό από την Αιθιοπία αποτελεί όντως μία εναλλακτική σύσταση. Η βασική ιδέα πίσω από την έννοια της εναλλακτικότητας είναι ότι τα ενδιαφέροντα ενός χρήστη γύρω από μία συγκεκριμένη κατηγορία αντικειμένων μπορεί να υποκρύπτουν ένα ενδιαφέρον για αντικείμενα άλλων κατηγοριών συγγενικών με τις κατηγορίες ενδιαφέροντος του χρήστη.

4. Αύξηση της ποικιλίας των συστάσεων: Τα συστήματα σύστασης τυπικά προτείνουν μία λίστα από τα πιο σχετικά αντικείμενα. Όταν όλα τα σχετικά αντικείμενα είναι πολύ όμοια μεταξύ τους, αυξάνεται ο κίνδυνος ο χρήστης τελικά να μην ενδιαφερθεί για κανένα από αυτά. Από την άλλη πλευρά, όταν η λίστα των συστάσεων αποτελείται από αντικείμενα διαφορετικών τύπων, υπάρχει μεγαλύτερη πιθανότητα ο χρήστης να ενδιαφερθεί τουλάχιστον για ένα από αυτά τα αντικείμενα. Παρέχοντας ποικιλία συστάσεων στο χρήστη εξασφαλίζεται το γεγονός ότι ο χρήστης δε θα βαρεθεί από την επαναλαμβανόμενη σύσταση παρόμοιων αντικειμένων.

Πέρα από αυτούς τους βασικούς στόχους, ένας αριθμούς ελάσσονος σημασίας στόχων μπορούν να καταγραφούν τόσο από την πλευρά τους χρήστη όσο και από την πλευρά του πωλητή. Από την πλευρά του χρήστη, τα συστήματα σύστασης μπορούν να βοηθήσουν στην αύξηση της ικανοποίησης από τη χρήση μίας ιστοσελίδας ή μία υπηρεσίας. Για παράδειγμα, ένας χρήστης που επανειλλημένα δέχεται σχετικές συστάσεις από το Amazon θα είναι πιο ευχαριστημένος από την εμπειρία του και θα έχει μεγαλύτερες πιθανότητες να χρησιμοποιήσει ξανά την ιστοσελίδα. Από την πλευρά του πωλητή, η διαδικασία της σύστασης μπορεί να παρέχει πληροφορίες για τις ανάγκες των χρηστών και να επιβοηθήσει στην προσπάθεια βελτίωσης της εμπειρίας χρήσης. Τέλος, η παροχή μίας εξήγησης για το λόγο για τον οποίο το αντικείμενο αυτό είναι σχετικό είναι επίσης χρήσιμο. Για παράδειγμα, στην περίπτωση του Netflix, οι συστάσεις παρέχονται μαζί με μία λίστα σχετικών ταινιών που έχουν παρακολουθηθεί στο παρελθόν.

Υπάρχει μία μεγάλη ποικιλία στα αντικείμενα που συστήνονται από τα διάφορα συστήματα σύστασης. Πολλά από αυτά δε συστήνουν προϊόντα, όπως για παράδειγμα το Facebook, το οποίο μπορεί να προτείνω κοινωνικές συνδέσεις μέσω των προτεινόμενων φίλων. Ακόμα και στην περίπτωση του Facebook, ωστόσο, απώτερος σκοπός είναι η αύξηση της χρησιμοποίησης της ιστοσελίδας και κατ' επέκταση των διαφημιστικών εσόδων.

1.3 Βασικά Μοντέλα Συστημάτων Σύστασης

Τα βασικά μοντέλα συστημάτων σύστασης δουλεύουν με δύο είδη δεδομένων, τα οποία είναι (i) οι αλληλεπιδράσεις χρηστών με τα αντικείμενα (πχ με

βαθμολογήσεις ή αγορές) και (ii) οι πληροφορίες για τα χαρακτηριστικά των χρηστών και των αντικειμένων, όπως για παράδειγμα συλλογή πληροφοριών για το προφίλ ενός χρήστη ή κατηγορίες στις οποίες ανήκουν τα αντικείμενα. Μέθοδοι οι οποίοι χρησιμοποιούν το πρώτο είδος αναφέρονται ως *μέθοδοι συνεργατικού φιλτραρίσματος*. Θα πρέπει να σημειωθεί ότι και τα συστήματα που χρησιμοποιούν το περιεχόμενο των αντικειμένων ως βασική συνιστώσα συστάσεων χρησιμοποιούν επίσης τον πίνακα βαθμολογιών, αλλά περιορίζονται συνήθως στις βαθμολογίες που έχει δώσει κάθε συγκεκριμένος χρήστης και όχι το σύνολο των χρηστών που έχουν χρησιμοποιήσει το σύστημα. Στα συστήματα *σύστασης βασισμένα σε γνώση*, οι συστάσεις βασίζονται σε συγκεκριμένες απαιτήσεις που έχουν ορίσει οι χρήστες. Αντί να χρησιμοποιούνται βάσεις δεδομένων με βαθμολογίες ή αγοραστικά δεδομένα, εξωτερικές βάσεις γνώσης και περιορισμοί χρησιμοποιούνται για τη δημιουργία των συστάσεων. Κάποια συστήματα σύστασης συνδυάζουν περισσότερες από μία από τις παραπάνω ιδέες και γι' αυτό αποκαλούνται *υβριδικά συστήματα*. Τα υβριδικά συστήματα μπορούν να συνδυάσουν τα θετικά στοιχεία διαφόρων συστημάτων σύστασης για να δημιουργήσουν τεχνικές που μπορούν να έχουν καλύτερα αποτελέσματα σε ένα ευρύ φάσμα περιπτώσεων.

1.3.1 Μοντέλα Συνεργατικού Φιλτραρίσματος

Τα μοντέλα συνεργατικού φιλτραρίσματος χρησιμοποιούν τη συνεργατική δύναμη των βαθμολογιών που παρέχονται από πολλαπλούς χρήστες για να κάνουν συστάσεις. Η μεγαλύτερη δυσκολία στη σχεδίαση μεθόδων συνεργατικού φιλτραρίσματος είναι το γεγονός ότι οι πίνακες των βαθμολογιών που λαμβάνουν ως είσοδο είναι στη συντριπτική πλειονότητα των περιπτώσεων αραιοί (sparse matrices). Ας σκεφτούμε ένα παράδειγμα όπου έχουμε ένα σύστημα ταινιών στο οποίο οι χρήστες μπορούν να δηλώσουν την προτίμησή τους ή μη απέναντι σε μία ταινία αφήνοντας μία βαθμολογία για κάθε ταινία. Εφόσον κάθε χρήστης έχει δει μόνο ένα πολύ μικρό κομμάτι του τεράστιου σύμπαντος των διαθέσιμων ταινιών που έχει παράγει το παγκόσμιο κινηματογραφικό σκηνικό, για κάθε χρήστη οι περισσότερες βαθμολογίες είναι κενές. Οι βαθμολογίες που δεν είναι κενές αναφέρονται και ως “παρατηρούμενες βαθμολογίες”.

Η βασική ιδέα πίσω από το συνεργατικό φιλτράρισμα είναι ότι οι κενές βαθμολογίες μπορούν να υπολογιστούν σε κάποιο βαθμό, επειδή οι παρατηρούμενες βαθμολογίες συνδέονται σε υψηλό βαθμό με βαθμολογίες άλλων χρηστών ή αντικειμένων. Για παράδειγμα, έστω δύο χρήστες, η Μαρία και ο Γιάννης, που έχουνε παρόμοια γούστα. Εάν για τα ίδια αντικείμενα, οι βαθμολογίες τους είναι παρόμοιες, τότε το σύστημα μπορεί να αντιληφθεί την ομοιότητα μεταξύ αυτών των δύο χρηστών. Σε αυτή την περίπτωση, αντικείμενα που έχουν βαθμολογηθεί μόνο από έναν από τους δύο έχουν μεγάλη πιθανότητα να λάβουν την ίδια βαθμολογία και από τον άλλο. Αυτή η ομοιότητα μπορεί να χρησιμοποιηθεί για να εξαχθούν συμπεράσματα για τις κενές βαθμολογίες. Τα περισσότερα συστήματα συνεργατικού φιλτραρίσματος χρησιμοποιούν την ομοιότητα που υπάρχει είτε μεταξύ αντικειμένων, είτε μεταξύ χρηστών. Ορισμένα μοντέλα χρησιμοποιούν και τους δύο τύπους συσχετίσεων. Επιπροσθέτως, μερικά μοντέλα χρησιμοποιούν προσεκτικά σχεδιασμένες τεχνικές βελτιστοποίησης για να δημιουργήσουν ένα μοντέλο εξάσκησης με τρόπο παρόμοιο με αυτόν που χρησιμοποιεί ένας ταξινομητής για να δημιουργήσει ένα μοντέλο εξάσκησης για δεδομένα που ανήκουν σε γνωστές κατηγορίες. Αυτό το μοντέλο χρησιμοποιείται εν συνεχεία για να εξάγει τις κενές βαθμολογίες. Υπάρχουν δύο τύποι μεθόδων που χρησιμοποιούνται για το συνεργατικό φιλτράρισμα, οι οποίοι αναφέρονται ως “μέθοδοι βασισμένες σε μνήμη” και “μέθοδοι βασισμένες σε μοντέλο”.

1. Μέθοδοι βασισμένες σε μνήμη: Οι μέθοδοι βασισμένες σε μνήμη αναφέρονται και ως αλγόριθμοι συνεργατικού φιτλραρίσματος γειτνίασης. Πρόκειται για τους πρώτες αλγορίθμους συνεργατικού φιλτραρίσματος στους οποίους οι βαθμολογίες των συνδυασμών χρήστη-αντικειμένου προβλέπονται με βάση τις περιοχές γειτνίασης. Αυτές οι περιοχές μπορούν να οριστούν με δύο τρόπους:
    - Συνεργατικό φιλτράρισμα με βάση τους χρήστες (user-user collaborative filtering): Σε αυτή την περίπτωση, οι βαθμολογίες που παρέχονται από χρήστες με παρόμοια ενδιαφέρονται με το χρήστη Α χρησιμοποιούνται για την εξαγωγή συστάσεων για το χρήστη Α. Έτσι, ο βασικός στόχος είναι να προσδιοριστούν χρήστες που είναι παρόμοιοι με το χρήστη Α και να προταθούν βαθμολογίες

που αντιστοιχούν στις κενές βαθμολογίες του χρήστη A υπολογίζοντας τους σταθμισμένους μέσους όρους των χρηστών που ανήκουν στην περιοχή γειτνίασης του A. Επιμένως, εάν ο Γιάννης και η Μαρία έχουν βαθμολογήσει ταινίες με παρόμοιο τρόπο στο παρελθόν, τότε μπορούν να χρησιμοποιηθούν οι παρατηρούμενες βαθμολογίες της Μαρίας για την ταινία “Star Wars Episode V” για να προβλεφθεί η κενή βαθμολογία του Γιάννη για τη συγκεκριμένη ταινία. Γενικά, οι k πιο κοντινοί χρήστες στο Γιάννη μπορούν να χρησιμοποιηθούν για να προκύψουν συμπεράσματα για τις κενές του βαθμολογίες. Συναρτήσεις ομοιότητας χρησιμοποιούνται επί των γραμμών του πίνακα βαθμολογιών για να προσδιοριστούν οι ομοιότητες μεταξύ των χρηστών.

- Συνεργατικό φιλτράρισμα με βάση τα αντικείμενα (item-item collaborative filtering): Προκειμένου να υπολογιστεί η βαθμολογία ενός αντικειμένου B από ένα χρήστη A απαιτείται σε πρώτη φάση ο προσδιορισμός ενός συνόλου S από αντικείμενα τα οποία είναι πιο σχετικά με το αντικείμενο B. Οι βαθμολογίες στο σύνολο αντικειμένων S που προσδιορίζονται από τον A χρησιμοποιούνται για να εντοπιστεί η προβλεπόμενη βαθμολογία του αντικειμένου B από τον A. Κατ' αυτό τον τρόπο, οι βαθμολογίες του Γιάννη για παρόμοιες ταινίες επιστημονικής φαντασίας, όπως το Alien και το Predator μπορούν να χρησιμοποιηθούν για την πρόβλεψη της βαθμολογίας του Γιάννη για την ταινία Terminator. Σε αυτή την περίπτωση, συναρτήσεις ομοιότητας χρησιμοποιούνται επί των στηλών του πίνακα βαθμολογιών για να προσδιοριστούν οι ομοιότητες μεταξύ των αντικειμένων.

Τα πλεονεκτήματων μεθόδων βασισμένων σε μνήμη είναι ότι είναι εύκολες να υλοποιηθούν και οι απορρέουσες συστάσεις είναι συνήθως έυκολο να εξηγηθούν. Από την άλλη πλευρά, οι μέθοδοι βασισμένοι σε μνήμη δε δουλεύουν αποδοτικά σε περιπτώσεις

αραιών πινάκων βαθμολογιών. Για παράδειγμα, μπορεί να είναι αρκετά δύσκολο να βρεθούν χρήστες παρόμοιοι με το Γιάννη που έχουν δώσει βαθμολογία για την ταινία *Inception*. Σε αυτές τις περιπτώσεις, είναι δύσκολο να προβλεφθεί με βεβαιότητα η βαθμολογία του Γιάννη για τη συγκεκριμένη ταινία. Με άλλα λόγια, τέτοιες μέθοδοι μπορεί να μην έχουν τη δυνατότητα να παράγουν αποδοτικά αποτελέσματα στο σύνολο των διαθέσιμων αντικειμένων. Παρόλα αυτά, η αδυναμία αυτή συχνά παραβλέπεται στις περιπτώσεις που απαιτούνται οι k-καλύτερες συστάσεις για κάθε χρήστη.

2. Μέθοδοι βασισμένες σε μοντέλα: Στις μεθόδους βασισμένες σε μοντέλα, μηχανική μάθηση και εξόρυξη δεδομένων χρησιμοποιούνται στα πλαίσια μοντέλων πρόβλεψης. Σε περιπτώσεις όπου το μοντέλο είναι παραμετροποιημένο, οι παράμετροι αυτού του μοντέλου “διδάσκονται” μέσω μίας δομής βελτιστοποίησης. Μερικά παραδείγματα τέτοιων μεθόδων βασισμένων σε μοντέλα περιλαμβάνουν τα δέντρα αποφάσεων, τα μοντέλα βασισμένα σε κανόνες, τις Μπαγιεσιανές μεθόδους και τα μοντέλα λανθάνοντα παράγοντα. Πολλές από αυτές τις μεθόδους όπως οι μέθοδοι λανθάνοντα παράγοντα έχουν ένα πολύ υψηλό επίπεδο κάλυψης ακόμα και για αραιούς πίνακες βαθμολογιών.

Παρόλο που οι αλγόριθμοι συνεργατικού φιλτραρίσματος βασισμένοι σε μνήμη εκτιμώνται για την απλότητά τους, τείνουν να είναι ευριστικοί στη φύση τους και δε λειτουργούν αποτελεσματικά σε όλες τις συνθήκες. Ωστόσο, η διάκριση ανάμεσα σε μεθόδους βασισμένες σε μνήμη και βασισμένες σε μοντέλα είναι κατά κάποιον τρόπο τεχνητή, εξαιτίας του γεγονότος ότι οι μέθοδοι βασισμένοι σε μνήμη μπορούν να θεωρηθούν ως ευριστικά μοντέλα βασισμένα σε ομοιότητα. Πρόσφατα, αποδείχτηκε ότι συνδυασμός των δύο παραπάνω μεθόδων μπορεί να επιφέρει αποτελέσματα πολύ υψηλής ακρίβειας.

1.3.2 Συστήματα Βασισμένα στο Περιεχόμενο

1 Εισαγωγή στα Συστήματα Σύστασης

Στα συτήματα σύστασης βασισμένα στο περιεχόμενο, τα περιγραφικά χαρακτηριστικά των αντικειμένων χρησιμοποιούνται για να προσδιοριστούν συστάσεις. Ο όρος “περιεχόμενο” αναεφέρεται σε αυτά τα χαρακτηριστικά. Στις μεθόδους βασισμένες στο περιεχόμενο, οι βαθμολογίες και η αγοραστική συμπεριφορά των χρηστών συνδυάζονται μαζί με τις πληροφορίες για το περιεχόμενο των διαθέσιμων αντικειμένων. Για παράδειγμα, ας θεωρήσουμε την περίπτωση όπου ο Γιάννης έχει δώσει υψηλή βαθμολογία στην ταινία Terminator, αλλά δεν έχουμε πρόσβαση στις βαθμολογίες των άλλων χρηστών. Επομένως, οι μέθοδοι συνεργατικού φιλτραρίσματος που εξετάστηκαν παραπάνω δεν μπορούν να χρησιμοποιηθούν. Παρόλα αυτά, η περιγραφή του αντικειμένου Terminator περιέχει παρόμοιες έννοιες που περιέχονται και σε άλλες ταινίες επιστημονικής φαντασίας όπως το Alien και το Predator. Σε αυτές τις περιπτώσεις, αυτές οι ταινίες προτείνονται στο Γιάννη.

Στις μεθόδους που βασίζονται σε περιεχόμενο, οι περιγραφές των αντικειμένων που περιέχουν και βαθμολογίες χρησιμοποιούνται ως δεδομένα εκπαίδευσης για την επίλυση ενός προβλήματος κατηγοριοποίησης ή μοντέλου παλινδρόμησης.

Για κάθε χρήστη, τα δεδομένα εκπαίδευσης αντιστοιχίζονται με τα αντικείμενα που αυτός ο χρήστης έχει αγοράσει ή βαθμολογήσει. Η μεταβλητή της κατηγορίας αντιστοιχίζεται σε συγκεκριμένες βαθμολογίες ή αγοραστικές συμπεριφορές. Αυτά τα δεδομένα χρησιμοποιούνται για τη δημιουργία ενός μοντέλο κατηγοριοποίησης ή παλινδρόμησης που είναι συγκεκριμένο για κάθε χρήστη. Αυτό το προσωποποιημένο μοντέλο χρησιμοποιείται για να προβλέψει εάν ο αντίστοιχος χρήστης θα ενδιαφερθεί για ένα αντικείμενο για το οποίο η βαθμολογία ή η αγοραστική του συμπεριφορά είναι άγνωστη.

Οι μέθοδοι βασισμένοι στο περιεχόμενο έχουν μερικά πλεονεκτήματα στην παραγωγή συστάσεων για καινούρια αντικείμενα, όταν επαρκή βαθμολογικά δεδομένα δεν είναι ακόμα διαθέσιμα για τα συγκεκριμένα αντικείμενα. Αυτό οφείλεται στο ότι άλλα αντικείμενα με παρόμοια χαρακτηριστικά μπορεί να έχουν ήδη βαθμολογηθεί από το χρήστη. Επομένως, το επιβλεπόμενο μοντέλο θα είναι σε θέση να προβλέψει τις βαθμολογίες με τη βοήθεια των χαρακτηριστικών των αντικειμένων και να δώσει συστάσεις ακόμα και όταν δεν υπάρχει ιστορικό βαθμολογιών για το συγκεκριμένο αντικείμενο.

Οι μέθοδοι βασισμένες σε περιεχόμενο έχουν και ορισμένα μειονεκτήματα:

1. Σε πολλές περιπτώσεις, οι μέθοδοι βασισμένες σε περιεχόμενο παρέχουν προφανείς συστάσεις εξαιτίας της χρήσης χαρακτηριστικών ή τετριμμένου περιεχομένου. Για παράδειγμα, εάν ένας χρήστης δεν έχει ποτέ ενδιαφερθεί για ένα αντικείμενο συγκεκριμένων κατηγοριών, το αντικείμενο αυτό δε θα προταθεί ποτέ στο συγκεκριμένο χρήστη. Αυτό οφείλεται στο ότι το κατασκευασμένο μοντέλο αναφέρεται αποκλειστικά στο συγκεκριμένο χρήστη και η γνώση που υπάρχει για τους υπόλοιπους χρήστες δεν αξιοποιείται. Αυτός το φαινόμενο τείνει να μειώσει την ποικιλία των προτεινόμενων αντικειμένων, γεγονός το οποίο δεν είναι επιθυμητό.

2. Παρόλο που οι μέθοδοι βασισμένες σε περιεχόμενο είναι αποτελεσματικές στο να παρέχουν συστάσεις για νέα αντικείμενα, δεν είναι το ίδιο αποτελεσματικές στο να προσφέρουν συστάσεις για νέους χρήστες. Αυτό οφείλεται στο ότι το μοντέλο εκπαίδευσης για κάθε χρήστη απαιτεί το ιστορικό των βαθμολογιών του. Στην πραγματικότητα είναι συνήθως σημαντικό να προϋπάρχει ένας μεγάλος αριθμός διαθέσιμων βαθμολογιών για το συγκεκριμένο χρήστη προκειμένου να έχουμε σωστές συστάσεις.

Επομένως, οι μέθοδοι βασισμένες στο περιεχόμενο έχουν διαφορετικά πλεονεκτήματα και μεινεκτήματα σε σχέση με τα συστήματα συνεργατικού φιλτραρίσματος.

Παρόλο που η παραπάνω περιγραφή παρέχει την τυπική θεώρηση των μεθόδων του είδους, μία ευρύτερη θεώρηση ενίοτε χρησιμοποιείται. Για παράδειγμα, οι χρήστες μπορούν να αναφέρουν συγκεκριμένες σχετικές λέξεις-κλειδιά στα προφίλ τους. Αυτά τα προφίλ μπορούν να συνδεθούν με περιγραφές αντικειμένων προκειμένου να πραγματοποιηθούν οι συστάσεις. Μία τέτοια προσέγγιση δε χρησιμοποιεί βαθμολογίες κατά τη διαδικασία των συστάσεων και είναι επομένως πολύ χρήσιμη σε σενάρια ψυχρής εκκίνησης. Παρόλα αυτά, τέτοιες μέθοδοι αναφέρονται κυρίως ως μία ξεχωριστή κατηγορία συστημάτων σύστασης που ονομάζονται Συστήματα Σύστασης βασισμένα σε Γνώση, καθώς οι μετρικές ομοιότητας βασίζονται στη γνώση του εκάστοτε πεδίου. Τα συστήματα σύστασης βασισμένα σε γνώση συχνά

θεωρείται ότι είναι άμεσα συνδεδεμένα με τα συστήματα σύστασης που βασίζονται στο περιεχόμενο, παρόλο που στις περισσότερες εκφάνσεις τους χρησιμοποιούν διαφορετικές προσεγγίσεις.

1.3.3 Συστήματα Σύστασης Βασισμένα σε Γνώση

Τόσο τα συστήματα που συνδέονται με περιεχόμενο όσο και τα συνεργατικά συστήματα απαιτούν ένα σημαντικό αριθμό δεδομένων σχετικών με παρελθοντικές αγοραστικές και βαθμολογικές εμπειρίες των χρηστών. Για παράδειγμα, τα συνεργατικά συστήματα απαιτούν την ύπαρξη ενός αρκετά πυκνού πίνακα βαθμολογήσεων για να πραγματοποιήσουν μελλοντικές συστάσεις. Το πρόβλημα αυτό αναφέρεται συχνά και ως πρόβλημα της ψυχρής εκκίνησης. Διαφορετικά συστήματα έχουν διαφορετικούς βαθμούς ευαισθησίας στο συγκεκεμένο πρόβλημα. Για παράδειγμα, τα συνεργατικά συστήαματα είναι τα πιο ευαίσθητα και δεν μπορούν να διαχειριστούν νέους χρήστες ή νέα αντικείμενα ικανοποιητικά. Τα συστήματα σύστασης βασισμένα στο περιεχόμενο είναι κάπως καλύτερα στη διαχείριση νέων αντικειμένων, αλλά ούτε και αυτά μπορούν να προσφέρουν ακριβείς συστάσεις σε νέους χρήστες.

Επιπλέον, αυτές οι μέθοδοι γενικά δεν είναι κατάλληλα διαμορφωμένες για πεδία ενδιαφέροντος όπου το προϊόν είναι μεγάλο βαθμό προσαρμόσιμο στις ανάγκες του χρήστη. Χαρακτηριστικά παραδείγματα αποτελούν τα ακίνητα, τα αυτοκίνητα, οι τουριστικές ανάγκες ή ακριβά προϊόντα πολυτελείας. Τέτοια αντικείμενα αγοράζονται σπάνια και αρκετές βαθμολογήσεις σπάνια είναι διαθέσιμες. Σε πολλές περιπτώσεις, το πεδίο των αντικειμένων μπορεί να είναι πολύπλοκο και μπορεί να υπάρχουν λίγες εμφανίσεις ενός αντικειμένου σε ένα συγκεκριμένο σύνολο κατηγοριών. Για παράδειγμα, κάποιος μπορεί να επιθυμεί να αγοράσει ένα ακίνητο με συγκεκριμένο αριθμό δωματίων, συγκεκριμένο γκαζόν, τοποθεσία κτλ. Εξαιτίας της πολυπλοκότητας περιγραφής του αντικειμένου, μπορεί να είναι δύσκολο να καταγραφεί ένα σαφές σύνολο από βαθμολογήσεις που αντανακλούν τις παρελθοντική σχέση ενός χρήστη με ένα συγκεκειμένο αντικείμενο. Παρόμοια, μία παλιά βαθμολογία σε ένα αυτοκίνητο με ένα συγκεκριμένο σύνολο επιλογών μπορεί να μην είναι σχετική με τα σημερινά δεδομένα.

Πώς μπορεί να γίνει διαχείριση τόσο μεγάλου βαθμού προσωποποίησης και ελαχιστότητας βαθμολογιών. Τα συστήματα σύστασης βασισμένα σε γνώση στηρίζονται εξ ολοκλήρου στη ρητή καταγραφή των απαιτήσεων των χρηστών για τα αντικείμενα ενδιαφέροντος. Παρόλα αυτά, σε τόσο πολύπλοκα πεδία ενδιαφέροντος είναι συχνά δύσκολο οι χρήστες να αντιληφθούν τον τρόπο με τον οποίο οι απαιτήσεις τους συνδέονται με τη διαθεσιμότητα του προϊόντος. Για παράδειγμα, ένας χρήστης μπορεί να μην έχει αντιληφθεί ότι ένα αυτοκίνητο με συγκεκριμένο συνδυασμό κατανάλωσης καυσίμου και ιπποδύναμης είναι διαθέσιμο με αποτέλεσμα να μην το αναζητήσει ποτέ. Επομένως, τέτοια συστήματα χρησιμοποιούν διαδραστική ανατροφοδότηση, η οποία επιτρέπει στη χρήστη να εξερευνήσει το εν δυνάμει πολύπλοκο πεδίο ενδιαφέροντος και να μάθει τα πλεονεκτήματα και τα μειονεκτήματα κάθε συνδυασμού που μπορεί να πραγματοποιηθεί. Η ανάκτηση και η η εξερεύνηση επιβοηθούνται εάν στην οντολογική περιγράφη εξετάζονται ως έννοιες όλα τα χαρακτηριστικά σε συνδυασμό με τα πλεονεκτήματα και τα μεινοκτήματά τους. Για παράδειγμα, εκτός από την έννοια “Αυτοκίνητο Με Μηχανή Αερίου” θα ήταν επιπλέον χρήσιμη η ύπαρξη της έννοιας “Οικολογικό Αυτοκίνητο”, αλλά και της έννοιας “Ακριβότερο Καύσιμο”.

Τα συστήματα σύστασης βασισμένα σε γνώση ενδείκνυνται ιδιαίτερα για την προώθηση αντικειμένων που δεν αγοράζονται ή προτιμώνται σε τακτική βάση. Επιπροσθέτως, στα αντικείμενα πεδίων που χρησιμοποιούνται συνήθως, οι χρήστες τείνουν να δηλώνουν πιο ρητά τις απαιτήσεις τους. Ένας χρήστης θα είναι πιο εύκολο να αποδεχτεί μία σύσταση για μία ταινία χωρίς πολλές ενστάσεις απ' ότι ένα σπίτι ή ένα αυτοκίνητο χωρίς να έχει ενδελεχείς πληροφορίες για το αντικείμενο και τη λογική πίσω από τη σύσταση. Επομένως, τα συστήματα σύστασης βασισμένα σε γνώση ταιριάζουν σε πεδία ενδιαφέροντος που είναι διαφορετικά από τα πεδία των συστημάτων συνεργατικού φιλτραρίσματος και βασιζόμενων στο περιεχόμενο. Γενικά, ενθαρρύνεται η χρήση των συστημάτων σύστασης βασισμένα σε γνώση στις παρακάτω περιπτώσεις:

1. Όταν οι χρήστες είναι διατεθειμένοι να δηλώσουν ξεκάθαρα τις απαιτήσεις τους. Σε αυτά τα πλαίσια, η διαδραστικότητα είναι βασικό κομμάτι αυτών των συστημάτων. Αξίζει να σημειωθεί ότι τα

συνεργατικά και βασιζόμενα στο περιεχόμενο συστήματα δεν επιτρέπουν αυτού του είδους τη λεπτομερή ανατροφοδότηση.

2. Όταν είναι δύσκολο να ληφθούν βαθμολογίες για ένα συγκεκριμένο τύπο αντικειμένου εξαιτίας της μεγάλης πολυπλοκότητας του πεδίου ενδιαφέροντος όσον αφορά τον τύπο αντικειμένων και τις διαθέσιμες επιλογές.

3. Όταν σε ορισμένα πεδία ενδιαφέροντος, όπως για παράδειγμα στον τομέα του hardware υπολογιστών, οι βαθμολογίες εξαρτώνται πολύ από το χρόνο. Οι βαθμολογίες για ένα παλιό αυτοκίνητο ή μία παλιά κάρτα γραφικών δεν είναι χρήσιμες, καθώς τέτοιου είδους προϊόντα δεν έχουν την ίδια αξία μερικά χρόνια μετά την κυκλοφορία τους.

Ένα πολύ σημαντικό κομμάτι των συστημάτων σύστασης βασισμένων σε γνώση είναι ο μεγαλύτερος έλεγχος που έχει ο χρήστης στον έλεγχο της κατεύθυνσης της διαδικασίας των συστάσεων. Αυτός ο μεγαλύτερος έλεγχος είναι άμεσο αποτέλεσμα της ανάγκης για ακριβή δήλωση των απαιτήσεών του σε ένα εν δυνάμει σύνθετο τομέα ενδιαφέροντος. Σε αυτά τα πλαίσια, η βασική διαφορά μεταξύ ενός συστήματος σύστασης βασισμένου σε γνώση και τών δύο προεξεταζόμενων συστημάτων είναι το γεγονός ότι τα άλλα συστήματα βασίζονται ως επί το πλείστον σε δεδομένα από το ιστορικό χρήσης της εφαρμογής, ενώ τα συστήματα σύστασης βασισμένα σε γνώση αξιοποιούν απευθείας πληροφορίες του χρήστη για τις επιθυμίες του. Ένα σημαντικό διαχωριστικό στοιχείο είναι η ελευθερία στην προσαρμογή του πεδίου ενδιαφέροντος. Τα συστήματα σύστασης βασισμένα σε γνώση επιτυγχάνουν εγγενώς αυτή την προσαρμοστικότητα είτε με τη δήλωση περιορισμών σε μία οντολογία, είτε με τη θέσπιση μέσων ομοιότητων που δρουν επί των αντικειμένων του πεδίο ενδιαφέροντος. Ορισμένα από αυτά τα συστήματα αξιοποιούν επιπλέον προσωπικά στοιχεία των χρηστών (για παράδειγμα δημογραφικά χαρακτηριστικά) σε συνδυασμό με τα χαρακτηριστικά των αντικειμένων ενδιαφέροντος, τα οποία λαμβάνονται από την αλληλεπίδραση του χρήστη με το σύστημα. Παρόλα αυτά, η πρακτική αυτή δεν είναι καθολική και δεν συμπεριλαμβάνεται σε όλα τα συστήματα σύστασης βασισμένα σε γνώση.

1 Εισαγωγή στα Συστήματα Σύστασης

Τα συστήματα σύστασης βασισμένα σε γνώση μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες με βάση τη μεθοδολογία αλληλεπίδρασης με το χρήστη και τη συνεπαγόμενη χρησιμοποίηση της βάσης γνώσης:

1. Συστήματα σύστασης βασισμένα σε περιορισμούς: Στα συστήματα σύστασης βασισμένα σε περιορισμούς, οι χρήστες τυπικά επιλέγουν απαιτήσεις ή περιορισμούς (πχ συγκεκριμένες τιμές) για τα χαρακτηριστικά των αντικειμένων. Επιπλέον, κανόνες που αναφέρονται στο συγκεκριμένο πεδίο χρησιμοποιούνται για να συνδέσουν απαιτήσεις των χρηστών με χαρακτηριστικά των αντικειμένων (για παράδειγμα η απαίτηση του χρήστη για την εύρεση αυτοκινήτων που έχουν αυτόματα πλοηγό αποκλείει αυτομάτως όλα τα αυτοκίνητα που έχουν κατασκευαστεί πριν το 2005). Ανάλογα με τον αριθμό και τον τύπο των επιστρεφόμενων αποτελεσμάτων, ο χρήστης μπορεί να έχει την ευκαιρία να αλλάξει τις αρχικές απαιτήσεις. Για παράδειγμα, μπορεί να χαλαρώσει κάποιους περιορισμούς, εάν επιστραφούν πολύ λίγες προτάσεις ή να προσθέσει επιπλέον εάν συμβαίνει το αντίθετο. Αυτή η διαδικασία αναζήτησης επαναλαμβάνεται εώς ότου ο χρήστης φτάσει στα επιθυμητά αποτελέσματα.

2. Συστήματα σύστασης βασισμένα σε περιπτώσεις: Στα συστήματα σύστασης βασισμένα σε περιπτώσεις, ορισμένες περιπτώσεις δηλώνονται από το χρήστη ως στόχοι. Μέτρα ομοιότητας ορίζονται ώστε να επιλεχθούν συγκεκριμένα χαρακτηριστικά και να συσταθούν τα κατάλληλα αντικείμενα. Τα μέτρα αυτά συνήθως είναι προσεκτικά επιλεγμένα με βάση το πεδίο ενδιαφέροντος. Επομένως, πρόκειται για διαφορετικά μέτρα ανάλογα με τη φύση του συστήματος και τους στόχους λειτουργίας. Τα επιστρεφόμενα αποτελέσματα χρησιμοποιούνται συνήθως ως νέοι στόχοι με κάποιες διαδραστικές διορθώσεις από το χρήστη για την παροχή των τελικών συστάσεων. Για παράδειγμα, όταν ένας χρήστης δει κάποιο επιστρεφόμενο αποτέλεσμα το οποίο σχεδόν ταιριάζει με αυτό που θέλει, μπορεί να επαναλάβει την αναζήτηση με αυτό το αντικείμενο ως στόχο και ενδεχομένως με κάποιες αλλαγές στα χαρακτηριστικά του κατά τις προτιμήσεις του. Αυτή η διαδραστική διαδικασία μπορεί να οδηγήσει πιο αποτελεσματικά το χρήστη στις επιθυμητές συστάσεις.

Πρέπει να σημειωθεί ότι και στις δύο περιπτώσεις, το σύστημα επιτρέπει στο χρήστη να αλλάξει τις απαιτήσεις που έχει ορίσει. Ωστόσο, ο τρόπος με τον οποίο αυτό γίνεται διαφέρει σε κάθε περίπτωση. Στα συστήματα σύστασης βασισμένα σε περιπτώσεις, παραδείγματα (ή περιπτώσεις) χρησιμοποιούνται ως σημεία αναφοράς χρησιμοποιούνται σε συνδυασμό με μέτρα ομοιότητας για να κατευθύνουν την αναζήτηση. Από την άλλη πλευρά, τα συστήματα βασισμένα σε περιορισμούς χρησιμοποιούν ρητά δηλωμένα χαρακτηριστικά για να πετύχουν τον ίδιο στόχο. Και στις δύο περίπτώσεις, τα παρουσιαζόμενα αποτελέσματα χρησιμοποιούνται για την αναζήτηση επιπρόσθετων επιλογών. Τα συστήματα σύστασης βασισμένα σε γνώση παίρνουν το όνομά τους από το γεγονός ότι κωδικοποιούν τη γνώση για το πεδίο ενδιαφέροντος σε περιορισμούς, κανόνες, μέτρα ομοιότητας και άλλα κατά τη διαδικασία της αναζήτησης. Για παράδειγμα, η σχεδίαση ενός μέτρου ομοιότητας ή ενός συγκεκριμένου περιορισμού απαιτείται γνώση σχετική με το πεδίο, η οποία είναι σημαντική για την ορθή λειτουργία του συστήματος. Γενικά, τα συστήματα σύστασης βασισμένα σε γνώση αντλούν πληροφορίες από σημαντικά ετερογενείς πηγές, σε αντίθεση με τα συνεργατικά συστήματα και τα συστήματα περιεχομένου, τα οποία λειτουργούν με βάση ομοιογενή αριθμητικά (κυρίως) δεδομένα από παρόμοιες πηγές. Σαν αποτέλεσμα, τα συστήματα σύστασης βασισμένα σε γνώση είναι σε μεγάλο βαθμό διαμορφώσιμα, αλλά δύσκολα γενικεύονται και αλλάζουν πεδίο ενδιαφέροντος με την ίδια αποτελεσματικότητα. Παρόλα αυτά, οι βασικές αρχές με τις οποίες δομούνται παραμένουν αμετάβλητες ανεξάρτητα από τη φύση και τα επιμέρους χαρακτηριστικά του πεδίου.

Η αλληλεπίδραση ανάμεσα στο χρήστη και το σύστημα μπορεί να πάρει τη μορφή ενός διαλογικού συστήματος, ενός συστήματος βασισμένου στην αναζήτηση ή ενός συστήματος κατεύθυνσης. Αυτές οι μορφές μπορεί να χρησιμοποιούνται από μόνες τους ή να συνδυάζονται σε κάθε επιμέρους σύστημα σύστασης βασισμένο σε γνώση:

1. Διαλογικά συστήματα: Σε αυτή την περίπτωση, οι προτιμήσεις του χρήστη ανιχνεύονται μέσω μίας επαναληπτικής διαδικασίας ανατροφοδότησης. Ο κύριος λόγος για την ύπαρξη αυτής της μορφής είναι η πολυπλοκότητα ορισμένων πεδίων, η οποία καθιστά την

καταγραφή των απαιτήσεων του χρήστη δυνατή μόνο στο περιβάλλον μίας στιαδιακής διαλογικής διαδικασίας.

2. Συστήματα βασισμένα σε αναζήτηση: Σε αυτά τα συστήματα, ο χρήστης καταδεικνύει τις απαιτήσεις, είτε μέσω ενός παραδείγματος είτε μέσα από την απάντηση σε μία σειρά ερωτήσεων του τύπου “Θα θέλατε να παρακολουθήσετε μία ταινία δράσης απόψε;”.

3. Συστήματα κατεύθυνσης: Στα συστήματα κατεύθυνσης, ο χρήστης έρχεται αρχικά αντιμέτωπος με κάποιες γενικές συστάσεις και στη συνέχεια καλείται να τις παραμετροποιήσει προκειμένου να καταλήξει στο ιδανικό επιθυμητό αποτέλεσμα. Μέσω μίας συνεχούς λήψης νέων αιτημάτων, το σύστημα είναι δυνατό να ανταπεξέλθει στις απαιτήσεις του χρήστη. Ένα σενάριο χρήσης αυτού του συστήματος θα εμφανιζόταν στην περίπτωση που το σύστημα έχει προτείνει την ταινία “Taken” ως ταινία δράσης στο χρήστη και ο χρήστης ανταποκρίνεται σε αυτή την πρόταση ζητώντας από το σύστημα “να προτείνει μία ταινία παρόμοια με το “Taken”, αλλά με πρωταγωνιστή τον Jason Statham”. Αυτά τα συστήματα σύστασης αναφέρονται και ως συστήματα σύστασης μέσω κριτικού φιλτραρίσματος.

Αυτές οι διαφορετικές μορφές ανταποκρίνονται σε διαφορετικούς τύπους συστημάτων σύστασης. Για παράδειγμα, τα συστήματα κριτικού φιλτραρίσματος είναι σχεδιασμένα για συστάσεις βασισμένες σε περιπτώσεις, καθώς οι χρήστες ασκούν κριτική σε συγκεκριμένα αποτελέσματα προκειμένου να φτάσουν στο επιθυμητό αποτέλεσμα. Από την άλλη πλευρά, ένα σύστημα βασισμένο σε αναζήτηση μπορεί να χρησιμοποιηθεί για την εύρεση των κατάλληλων περιορισμών σε συστήματα σύστασης βασισμένα σε περιορισμούς. Μερικές μορφές κατεύθυνσης μπορούν να χρησιμοποιηθούν τόσο με τα συστήματα βασισμένα σε περιορισμούς όσο και σε αυτά που είναι βασισμένα σε περιπτώσεις. Επιπροσθέτως, διαφορετικές μορφές κατεύθυνσης μπορούν να χρησιμοποιηθούν σε συνδυασμό σε ένα σύστημα βασισμένο σε γνώση. Δεν υπάρχουν αυτστηροί κανόνες όσον αφορά τον τρόπο με τον οποίο μπορεί ένα σχεδιαστεί μία διεπαφή για ένα σύστημα σύστασης. Είναι στόχος είναι πάντα να κατευθυνθεί ο χρήστης μετά από κάθε χρήση του συστήματος σε επιθυμητά και ενδιαφέροντα αποτελέσματα.

Τυπικά παραδείγμα της διαδραστικής διαδικασίας στα συστήματα βασισμένα σε περιορισμούς και περιπτώσεις φαίνονται παρακάτω. Η συνολική διαδραστική προσέγγιση είναι αρκετά παρόμοια. Η βασική διαφορά εμφανίζεται όσον αφορά τον τρόπο με τον οποίο ο χρήστης δηλώνει τα ερωτήματα και αλληλεπιδρά με το σύστημα για περαιτέρω επεξεργασία των αποτελεσμάτων. Στα συστήματα βασισμένα σε περιορισμούς, συγκεκριμένες απαιτήσεις (ή περιορισμοί) σημειώνονται από το χρήστη, ενώ στα συστήματα βασισμένα σε περιπτώσεις, συγκεκριμένοι στόχοι (ή περιπτώσεις) σημειώνονται από το χρήστη. Κατ' αναλογία, διαφορετικές μέθοδοι διαδραστικών διαδικασιών χρησιμοποιούνται από τα δύο συστήματα. Πιο συγκεκριμένα, στα συστήματα βασισμένα σε περιορισμούς η αρχική αναζήτηση διαφοροποιείται μέσω των διαδικασιών της προσθήκης, της διαγραφής και της αλλαγής των αρχικών δηλωμένων απαιτήσεων. Στα συστήματα που βασίζονται σε περιπτώσεις, είτε ο χρήστης αλλάζει τον αρχικό στόχο-κατάσταση, είτε τα αποτελέσματα περιορίζονται με βάση διαδικασίες κατευθυνόμενης κριτικής. Σε αυτές τις διαδικασίες, ο χρήστης απλώς δηλώνει εάν ένα συγκεκριμένο χαρακτηριστικό στα αποτελέσματα της έρευνας χρειάζεται να αυξηθεί, να μειωθεί ή να αλλάξει με συγκεκριμένο τρόπο. Αυτή η προσέγγιση αντιπροσωπεύει έναν πιο διαλογικό τρόπο από την απλή αλλαγή των απαιτήσεων που περιγράφηκε παραπάνω. Και στις δύο περιπτώσεις, ωστόσο ο κοινός στόχος είναι να βρίσκονται στη θέση είτε άμεσα, είτε έμμεσα να δηλώνουν ακριβώς τις επιθυμητές τους απαιτήσεις. Στα συστήματα βασισμένα σε περιορισμούς, το πρόβλημα αυτό λύνεται εν μέρει μέσω της χρήσης κανόνων βασισμένων σε γνώση, οι οποίοι αντιστοιχίζουν απαιτήσεις των χρηστών με χαρακτηριστικά προϊόντων. Στα συστήματα βασισμένα σε περιπτώσεις, το πρόβλημα αυτό αντιμετωπίζεται μέσα από μία διαλογική διαδικασία κριτικής. Ο διαδραστικός χαρακτήρας είναι κοινός και στα δύο συστήματα και είναι σημαντικός στην επιβοήθηση του χρήστη να αναζητήσει αντικείμενα που ικανοποιούν τις ανάγκες του σε πολύπλοκα πεδία ενδιαφέροντος.

Είναι αξιοσημείωτο ότι οι περισσότερες μορφές συστημάτων σύστασης βασισμένων σε γνώση εξαρτώνται σε μεγάλο βαθμό από τις περιγραφές των αντικειμένων, ο οποίες αντιμετωπίζονται με σημασιολογική χροιά και όχι με λεξικογραφική, όπως γίνεται στα συστήματα βασισμένα στο περιεχόμενο. Αυτή είναι μία φυσική συνέπεια της εγγενούς πολυπλοκότητας των

συστημάτων σύστασης βασισμένων σε γνώση στα οποία η γνώση σχετική με το πεδίο μπορεί πιο εύκολα να κωδικοποιηθεί σε σχεσιακά χαρακτηριστικά. Για παράδειγμα, τα δεδομένα για κάποιες ταινίες αναφέρονται στον παρακάτω πίνακα όπως αυτές αναφέρονται σε μία βάση δεδομένων.

ID	Τίτλος	Είδη	Πρωταγωνιστές
1	Snatch	Comedy, Crime	Jason Statham, Brad Pitt, Benicio Del Toro
2	Two Smoking Barrels	Comedy, Crime	Jason Flemyng, Dexter Fletcher, Nick Moran
3	The Usual Suspects	Crime, Drama, Mystery	Kevin Spacey, Gabriel Byrne
4	Back to the Future	Adventure, Comedy, Sci-Fi	Michael J. Fox, Christopher Lloyd
5	Star Trek	Action, Adventure, Sci-Fi	Chris Pine, Zachary Quinto, Simon Pegg

Στην περίπτωση των συστημάτων βασισμένων σε περιπτώσεις, οι μετρικές ομοιότητας καθορίζονται στα πλαίσια αυτών των χαρακτηριστικών ώστε να παρέχουν παρόμοιες ταινίες σύμφωνα με μία ή περισσότερες ταινίες που έχει ορίσει ο χρήστης. Στα συστήματα βασισμένα σε περιορισμούς, τα ερωτήματα προσδιορίζονται με βάση την επιλογή συγκεκριμένων χαρακτηριστικών, πχ Comedy Movies ή Movies Starring Jason Statham. Στα συστήματα βασισμένα σε περιπτώσεις, χρήστης θα επέλγε πχ την ταινία Snatch και με βάση αυτή θα γίνονταν οι συστάσεις μίας ή περισσότερων άλλων ταινιών από τις παραπάνω.

1.3.3.1 Συστήματα βασισμένα σε περιορισμούς

Τα συστήματα βασισμένα σε περιορισμούς επιτρέπουν στο χρήστη να αναφερθεί σε συγκεκριμένες απαιτήσεις ή περιορισμούν όσον αφορά τα χαρακτηριστικά των αντικειμένων. Επιπλέον, ένα σύνολο κανόνων χρησιμοποιείται προκειμένου να αντιστοιχιθούν οι απαιτήσεις του χρήστη με τα χαρακτηριστικά των αντικειμένων. Παρόλα αυτά, οι χρήστες μπορεί να μη δηλώνουν πάντα τα ερωτήματά τους με βάση τα χαρακτηριστικά που περιγράφουν τα αντικείμενα. Επομένως, ένα επιπλέον σύνολο κανόνων απαιτείται για να συσχετιστούν οι απαιτήσεις του χρήστη με τα χαρακτηριστικά των προϊόντων. Αν είχαμε για παράδειγμα ένα σύστημα

σύστασης καταλληλότητας σπιτιών κάποια χαρακτηριστικά θα μπορούσαν να είναι τα παρακάτω:

Οικογενειακή Κατάσταση (επιλογή από κατηγορίες), *μέγεθος οικογένειας* (αριθμητική τιμή), *προτίμηση για αστική ή προαστική οικία* (δυαδική τιμή), *αριθμός κρεβατοκάμαρων* (αριθμητική τιμή), *διαθέσιμα χρήματα* (αριθμητική τιμή)

Αυτά τα χαρακτηριστικά μπορεί να αντιπροσωπεύουν είτε ιδιότητες του χρήστη (πχ τα δημογραφικά χαρακτηριστικά που αναφέρονται παραπάνω), είτε απαιτήσεις που αφορούν το προϊόν. Τέτοιες απαιτήσεις συνήθως σημειώνονται διαδραστικά κατά τη διάρκεια του διαλόγου ανάμεσα στο χρήστη και το σύστημα σύστασης. Η μετατροπή μερικών από τις απαιτήσεις αυτές, όπως για παράδειγμα αυτή της μέγιστης τιμής είναι εύκολο να αντιστοιχιθεί σε χαρακτηριστικά των αντικειμένων μπορεί να είναι εύκολη, ωστόσο άλλες μετατροπές, όπως για παράδειγμα αυτή της προτίμησης για αστική ή προαστική κατοικία μπορεί να μην είναι τόσο προφανείς. Σε αυτές τις μη προφανείς περιπτώσεις, το ρόλο του διερμηνευτή των απαιτήσεων του χρήστη αναλαμβάνουν οι βάσεις γνώσεις με τις οποίες είναι εφοδιασμένα αυτά τα συστήματα. Για παράδειγμα, η προτίμηση του χρήστη για αστική μία προαστική οικία μπορεί να ερμηνευθεί μέσω μία σειράς κανόνων που συνδέουν σπίτια με περιοχές και περιοχές με το χαρακτηριστικό “αστική περιοχή” ή με το χαρακτηριστικό “προάστιο”.

Τέτοιοι κανόνες αναφέρονται συχνά και ως περιπτώσεις φιλτραρίσματος επειδή με την αντιστοίχηση των απαιτήσεων του χρήστη σε περιορισμούς φιλτράρουν τα αποτελέσματα που πρόκειται να παρουσιαστούν από το σύστημα. Αξίζει να σημειωθεί ότι αυτοί οι τύποι κανόνων προκύπτουν είτε από το πεδίο των προϊόντων, είτε, πιο σπάνια μπορεί να προκύπτουν από ιστορική εξόρυξη τέτοιων συνόλων. Στη συγκεκριμένη περίπτωση, είναι προφανές ότι ο κανόνας αυτός μπορεί να προκύψει χρησιμοποιώντας τη δημόσια γεωγραφική πληροφορία που έχουμε για κάθε περιοχή. Άλλο παράδειγμα είναι το πεδίο της αυτοκίνησης, όπου ορισμένες επιπλέον δυνατότητες μπορούν να γίνουν διαθέσιμες μόνο αν το αυτοκίνητο έχει συγκεκριμένα χαρακτηριστικά. Για παράδειγμα, μία μηχανή υψηλών οκτανίων μπορεί να είναι διαθέσιμη μόνο σε ένα αγωνιστικό μοντέλο. Τέτοιες συνθήκες αναφέρονται και ως συνθήκες συμβατότητας, επειδή μπορούν να αξιοποιηθούν για τη γρήγορη ανακάλυψη

ασυνεπειών μεταξύ των απαιτήσεων του χρήστη και των συνδυασμών που πραγματοποιούνται στο πεδίο ενδιαφέροντος. Υπό αυτό το πρίσμα, η ιστοσελίδα Edmunds.com δεν επιτρέπει στους χρήστες να δηλώσουν ασυνεπής απαιτήσεις χάρη στον περιορισμό των επιλογών μετά από κάθε δήλωση απαίτησης από το χρήστη. Σε άλλες περιπτώσεις όπου δεν είναι δυνατός ο έλεγχος για ασυνέπειες, αυτές μπορούν να γίνουν αντιληπτές κατά τη διάρκεια επεξεργασίας του αιτήματος του χρήστη και να δοθεί το κενό σύνολο αντικειμένων ως απάντηση στο ασυνεπές ερώτημα.

Άλλες συνθήκες ασυμβατότητας μπορεί να αναφέρονται στα ίδια τα χαρακτηριστικά του χρήστη, ειδικά στην περίπτωση που ο χρήστης καταθέτει προσωπικά του στοιχεία πριν από τη διαδικασία των συστάσεων. Για παράδειγμα, στην εφαρμογή για εύρεση κατάλληλου σπιτιού που περιγράφηκε παραπάνω μπορεί να υπάρχει επιπλέον ο κανόνας:

$$Marital.Status = single \Rightarrow Min.Bedrooms \leq 5$$

Σύμφωνα με τον κανόνα αυτό, οι άγαμοι χρήστες προτιμούν να αγοράζουν ή να νοικιάζουν σπίτια με λιγότερα από πέντε υπνοδωμάτια. Τέτοιοι κανόνες προκύπτουν από την ανάλυση του πεδίου ενδιαφέροντος (που έχει γίνει κατά τη δημιουργία της μηχανής συστάσεων), διευκολύνουν το φιλτράρισμα των αποτελεσμάτων και σε γενικές γραμμές βοηθάνε στην παροχή συστάσεων μεγαλύτερης ακρίβειας. Ομοίως ισχύει ο κανόνας:

$$Family.Size \geq 5 \Rightarrow Min.Bedrooms \geq 3$$

σύμφωνα με τον οποίο οι πολυμελείς οικογένειες προτιμούν σπίτια με τουλάχιστον τρεις κρεβατοκάμαρες. Παρότι χρήσιμοι στις περισσότερες περιπτώσεις, τέτοιοι κανόνες συχνά δημιουργούν συνθήκες ασυμβατότητας στο σύστημα και χρήζουν ειδικής αντιμετώπισης.

Σύμφωνα με όλα τα παραπάνω, υπάρχουν τρεις πρωταρχικοί τύποι εισόδου στα συστήματα σύστασης βασισμένα σε περιορισμούς:

1. Η πρώτη κατηγορία δεδομένων εισόδου αναφέρεται στα εγγενή χαρακτηριστικά του χρήστη (πχ δημογραφικά χαρακτηριστικά, επιθυμία ρίσκου κλπ) και συγκεκριμένες απαιτήσεις για το προϊόν (πχ ελάχιστος αριθμός υπνοδωματίων). Κάποιοι από αυτούς τους περιορισμούς είναι εύκολο να αντιστοιχηθούνε με τα χαρακτηριστικά του προϊόντος, ενώ

άλλοι πρέπει να αντιστοιχηθούνε μέσω δεδομένων που είναι αποθηκευμένα σε μια βάση γνώσης. Στις περισσότερες περιπτώσεις, τα χαρακτηριστικά του χρήστη και οι απαιτήσεις του δηλώνονται διαδραστικά για κάθε διαδικασία σύστασης ξεχωριστά και δεν αποθηκεύονται για μελλοντικές χρήσεις. Επομένως, αν ένας άλλος χρήστης χρησιμοποιήσει το ίδιο σύνολο απαιτήσεων και χαρακτηριστικών, θα λάβει τα ίδια αποτελέσματα. Σε αυτό το σημείο φαίνεται και η καίρια διαφορά των συστημάτων σύστασης βασισμένων σε γνώση σε σχέση με τους άλλους τύπους συστημάτων σύστασης, όπου η εξατομίκευση των συστάσεων βασίζεται σε δεδομένα που έχουν εκμαιευθεί από προηγούμενες χρήσεις του συστήματος.

2. Η δεύτερη κατηγορία δεδομένων εισόδου αναπαρίσταται από τις βάσεις γνώσης, οι οποίες συνδέουν τα χαρακτηριστικά και τις απαιτήσεις των χρηστών με χαρακτηριστικά των προϊόντων. Η σύνδεση αυτή μπορεί να γίνει, είτε άμεσα, είτε έμμεσα:

   - Άμεσα: Αυτοί οι κανόνες συνδέουν απαιτήσεις των χρηστών με απαραίτητα χαρακτηριστικά των προϊόντων. Ένα παράδειγμα τέτοιων κανόνων είναι το ακόλουθο:

     $$Suburban.Or.Rural = Suburban \Rightarrow Locality = (List\ of\ releveant\ Localities)$$

     $$Min/Bedrooms \geq 3 \Rightarrow 100.000$$

     Τέτοιοι κανόνες αναφέρονται και ως κανόνες φιλτραρίσματος.

   - Έμμεσα: Αυτοί οι κανόνες συνδέουν χαρακτηριστικά των χρηστών με αναμενόμενα, αλλά όχι απαραίτητα χαρακτηριστικά των προϊόντων. Επομένως, τέτοιοι κανόνες μπορού να θεωρηθούν και ως ένας έμμεσος τρόπος για τη συσχέτιση απαιτήσεων των χρηστών με περιορισμούς των προϊόντων. Παραδείγματα τέτοιων κανόνων είναι:

     $$Family.Size \geq 5 \Rightarrow Min.Bedrooms \geq 3$$

     $$Family.Size \geq 5 \Rightarrow Min.Bathrooms \geq 2$$

     Αξίζει να σημειωθεί ότι οι συνθήκες και στις δύο πλευρές των κανόνων αντιπροσωπεύουν χαρακτηριστικά των χρηστών τα

οποία μπορούν να συνδεθούν με χαρακτηριστικά των προϊόντων εύκολα. Αυτοί οι περιορισμοί αντιπροσωπεύουν περιορισμούς συμβατότητας. Στην περίπτωση που οι περιορισμοί συμβαότητας ή οι κανόνες φιλτραρίσματος είναι ασυνεπείς με τους περιορισμούς που έχει ορίσει ο χρήστης το σύνολο των προτεινόμενων συστάσεων είναι κενό.

Οι προαναφερθείσες βάσεις γνώσεων προκύπτουν από δημόσια διαθέσιμη πληροφορία, εμπειρογνώμονες, παλαιότερη εμπειρία ή εξόρυξη δεδομένων από διαθέσιμα σύνολα δεδομένων. Επομένως, σημαντικός μόχθος χρειάζεται για τη δημιουργία μίας βάσης γνώσης.

3. Τέλος, ο κατάλογος προϊόντων περιέχει μία λίστα από όλα τα προϊόντα μαζί με τα χαρακτηριστικά τους. Για παράδειγμα, τα χαρακτηριστικά των ταινιών που αναγράφονται παραπάνω είναι τα είδη τους και οι ηθοποιοί που πρωταγωνιστούν.

Επομένως, το πρόβλημα της σύστασης στη συγκεκριμένη περίπτωση μετασχηματίζεται στην προσπάθεια εύρεσης όλων των διαθέσιμων προϊόντων από τον κατόλογο που ικανοποιούν τις απαιτήσεις του χρήστη και τους κανόνες που έχουν οριστεί στη βάση γνώσης.

1.3.3.2 Επιστρέφοντας σχετικά αποτελέσματα

Το πρόβλημα της επιστροφής σχετικών αποτελεσμάτων μπορεί να μελετηθεί ως μία εκδοχή του προβλήματος ικανοποίησης περιορισμούς θεωρώντας κάθε αντικείμενο του καταλόγου ως έναν περιορισμό των χαρακτηριστικών και εκφράζοντας έτσι τον κατάλογο σε κανονική διαζευτική μορφή. Αυτή η έκφραση στη συνέχεια συνδυάζεται με τους κανόνες στη βάση γνώσης για να καθοριστεί εάν μία από κοινού συνεπής περιοχή του καταλόγου είναι δυνατή.

Πιο απλά, ένα σύνολο κανόνων και απαιτήσεων μπορεί να θεωρηθεί ως μία διαδικασία φιλτραρίσματος δεδομένων στον κατάλογο. Όλες οι απαιτήσεις του χρήστη είναι οι ενεργοί κανόνες σχετικοί με το χρήστη και χρησιμοποιούνται για τη δημιουργία ενός αιτήματος στη βάση δεδομένων των αντικειμένων του πεδίου ενδιαφέροντος. Τα βήματα για τη δημιουργία ενός τέτοιου αιτήματος φιλτραρίσματος είναι τα ακόλουθα:

1. Για κάθε απαίτηση (ή προσωπικό στοιχείο) που έχει δηλώσει ο χρήστης ελέγχεται εάν κάποια από αυτές ενεργοποιεί κάποιο κανόνα στη βάση γνώσης. Εάν κάτι τέτοιο ισχύει, τότε, τα αποτελέσματα των κανόνων προστίθενται στις απαιτήσεις που έχει ήδη ορίσει ο χρήστης. Η διαδικασία επαναλαμβάνεται ελέγχοντας κάθε φορά τις νέες απαιτήσεις έως ότου να μην ενεργοποιείται κανένας επιπλέον κανόνας της βάσης γνώσης. Για παράδειγμα, στην περίπτωση του συστήματος εύρεσης κατάλληλου σπιτιού, εάν ο χρήστης έχει ορίσει ως απαιτήσεις/προσωπικά στοιχεία ότι η οικογένειά του απαρτίζεται από 6 άτομα (Family.Size = 6) και στη βάση γνώσης υπάρχουν οι κανόνες:

$$Family.Size \geq 5 \Rightarrow Min.Bedrooms \geq 3$$
$$Family.Size \geq 5 \Rightarrow Min.Bathrooms \geq 2$$
$$Min.Bedrooms \geq 3 \Rightarrow Price \geq 100.000$$
$$Min.Bedrooms \geq 3 \Rightarrow Bedrooms \geq 3$$
$$Min.Bathrooms \geq 2 \Rightarrow Bathrooms \geq 2$$

   τότε, κατά την εκτέλεση αυτού του βήματος, θα προστεθούν αρχικά στις απαιτήσεις οι δύο πρώτοι κανόνες, ενώ σε επόμενο βήμα θα προστεθούν και οι άλλοι τρεις.

2. Αυτές οι επιπλέον απαιτήσεις χρησιμοποιούνται για τη δημιουργία ενός αιτήματος στη βάση δεδομένων σε κανονική συζευτική μορφή. Πρόκειται για ένα κλασσικό ερώτημα σε βάση δεδομένων, το οποίο υπολογίζει την τομή των παρακάτω περιορισμών, όπως αυτή εμφανίζεται στον κατάλογο προϊόντων:

$$(Bedrooms) \geq 3 \wedge (Bathrooms \geq 2) \wedge (Price \geq 100.000)$$

3. Το ερώτημα που έχει υπολογιστεί από την παραπάνω διαδικασία χρησιμοποιείται για τη λήψη των αντικειμένων του καταλόγου που ικανοποιούν όλες τις απαιτήσεις του χρήστη.

Σε αυτό το σημείο πρέπει να τονιστεί ότι μετά την εκτέλεση του ερωτήματος, τα δεδομένα του χρήστη δεν αποθηκεύεται στη μηχανή συστάσεων (η αποθήκευση των δεδομένων μέσω cookies από την αντίστοιχη ιστοσελίδα αποτελεί τελείως διαφορετικό ζήτημα) και σε αυτά τα πλαίσια, η πληροφορία που παρείχε ο χρήστης χάνεται μετά την εκτέλεση. Κατά την παρουσίαση των αποτελεσμάτων, ο χρήστης μπορεί να συνεχίσει να παραμετροποιεί τις

απαιτήσεις του και να ανακαλύψει συστάσεις που μπορεί στην αρχή της αναζήτησης να μην είχε υπόψη του ότι ικανοποιούν τις ανάγκες του.

1.3.3.3 Συστήματα Σύστασης Βασισμένα σε Περιπτώσεις

Στην περίπτωση των συστημάτων βασισμένων σε περιπτώσεις, χρησιμοποιούνται μετρικές ομοιότητας για να ανακτηθούν παραδείγματα που είναι παρόμοια με τις περιπτώσεις που έχει θέσει ο χρήστης. Για παράδειγμα, στην περίπτωση ενός συστήματος σύστασης που αναφέρεται στον τομέα του κινηματογράφου, ο χρήστης αναφέρει παραδείγματα ταινιών που του άρεσαν προκειμένου το σύστημα να παρέχει παρόμοιες ταινίες που έχουν υψηλή πιθανότητα να αρέσουν επίσης στο χρήστη. Εάν καμία άλλη ταινία δεν ανταποκρίνεται επακριβώς στα χαρακτηριστικά της ταινίας που δήλωσε ο χρήστης, τότε τα πλαίσια της αναζήτησης χαλαρώνουν και επιδιώκεται η παρουσίαση των όσο το δυνατόν πιο σχετικών ταινιών στο χρήστη. Επομένως, αντίθετα με τα συστήματα σύστασης βασισμένα σε περιορισμούς, στα συστήματα σύστασης βασισμένα σε περιπτώσεις η επιστροφή του κενού συνόλου ως σύνολο συστάσεων δεν αποτελεί πρόβλημα.

Υπάρχουν επίσης ουσιαστικές διαφορές ανάμεσα στα συστήματα σύστασης βασισμένα σε περιορισμούς και τα συστήματα βασισμένα σε περιπτώσεις όσον αφορά τον τρόπο με τον οποίο τα αρχικά αποτελέσματα αξιοποιούνται από το χρήστη. Τα συστήματα σύστασης βασισμένα σε περιορισμούς χρησιμοποιούν τη χαλάρωση, την τροποποίηση, και την ενίσχυση των απαιτήσεων για να προσφέρουν στο χρήστη εργαλεία εργασίας επί των αποτελεσμάτων. Τα πρώτα συστήματα σύστασης βασιμένα σε περιπτώσεις χρησιμοποίησαν ως βασική τακτική την επαναλαμβανόμενη αλλαγή του ερωτήματος του χρήστη έως ότου βρεθεί ένα επιθυμητό αποτέλεσμα. Η συγκεκριμένη διαδικασία ήταν χρονοβόρα και ζημίωνε την εμπειρία του χρήστη, γεγονός το οποίο οδήγησε στην ανάπτυξη της τεχνικής της κριτικής των αποτελεσμάτων. Η γενική ιδέα της κριτικής των αποτελεσμάτων είναι η επιλογή από το χρήστη ενός ή περισσότερων αποτελεσμάτων και η εκτέλεση επιπλέον ερωτημάτων με την ακόλουθη μορφή:

“Εμφάνισε περισσότερα αποτελέσματα σαν το X, αλλά με χαρακτηριστικά Y που είναι διαφορετικά από το αρχικό αντικείμενο Z που επέλεξα”

Μία σημαντική διαφοροποίηση της μεθόδου υπάρχει ως προς τον αν ένα ή περισσότερα χαρακτηριστικά μπορούν να επιλεγούν για επανεξέταση μετά από κάθε ερώτημα και ως προς τον τρόπο με τον οποίο πραγματοποιείται αυτή η επανεξέταση. Ο κύριος στόχος της κριτικής των αποτελεσμάτων είναι η υποστήριξη μίας διαδραστικής μετακίνησης του χρήστη μέσα στο πεδίο των αποτελεσμάτων προκειμένου να αντιληφθεί σταδιακά τις επιλογές που είναι διαθέσιμες στο πεδίο ενδιαφέροντος. Είναι συχνά δυνατό μέσω της επαναλαμβανόμενης διαδραστικής εξερεύνησης, ο χρήστης να ανακαλύψει αντικείμενα τα οποία δε θα μπορούσε μέσω της αρχικής διαδικασίας συστάσεων.

Για παράδειγμα, ας θεωρήσουμε την περίπτωση ενός συστήματος σύστασης στο οποίο ο χρήστης αναζητά ταινίες που μπορεί να τον ενδιαφέρουν με βάση παρόμοιες ταινίες που έχει δει. Ο χρήστης μπορεί, είτε να περιγράψει τη συγκεκριμένη ταινία συμπληρώνοντας το είδος, τους ηθοποιούς και άλλα στοιχεία για την ταινία ή να δηλώσει το ίδιο το όνομα της ταινίας. Η πρώτη περιγραφή ταιριάζει καλύτερα στις περιπτώσεις που ο χρήστης δε θυμάται ονόματα ταινιών που του αρέσουν ή θέλει να κάνει μία πολύ συγκεκριμένη περιγραφή ταινίας, ενώ η δεύτερη είναι πιο άμεση, ευκολότερη στην κατανόηση και βοηθάει ιδιαίτερα όταν το πεδίο ενδιαφέροντος είναι πολύ πολύπλοκο ώστε αντικείμενά του να περιγραφούν ως προς τα χαρακτηριστικά τους από το χρήστη. Μία περίπτωση πολύπλοκου πεδίου ενδιαφέροντοςε ίναι το πεδίο των ψηφιακών καμερών, όπου είναι δύσκολο ένας μη ειδικός να περιγράψει επακριβώς τα χαρακτηριστικά μίας ψηφιακής κάμερας που τον ενδιαφέρει. Παρόλα αυτά, είναι πολύ εύκολο για τον οποιονδήποτε να χρησιμοποιήσει το μοντέλο της φωτογραφικής κάμερας του φίλου του ως παράδειγμα και εφαλτήριο για την εκκίνηση της διαδικασίας των συστάσεων. Παραδείγματα τέτοιων διαδραστικών εφαρμογών εμφανίζονται παρακάτω.

Το σύστημα χρησιμοποιεί το ερώτημα που έχει θέσει ο χρήστης σε συνδυασμό με συναρτήσεις ομοιότητας ώστε να παράγει επιθυμητά αποτελέσματα. Σε τελικό στάδιο, μετά την εμφάνιση των αποτελεσμάτων ο χρήστης μπορεί να ενδιαφερθεί για κάποια από τις ταινίες που του προτάθηκαν, αλλά για κάποιο λόγο η συγκεκριμένη ταινία δεν ικανοποιεί τους στόχους του (πχ επειδή έχει ήδη δει αυτή την ταινία). Σε αυτό το σημείο ο χρήστης μπορεί να χρησιμοποιήσει τη συγκεκριμένη ταινία ως σημείο έναρξης και να επαναλάβει

την αναζήτηση ή ακόμα και να συνδυάσει τις δύο ταινίες (την πρώτη που παρείχε ως παράδειγμα και αυτή τη δεύτερη) προκειμένου να μπορέσει να μετακινηθεί σε καλύτερα αποτελέσματα. Θα πρέπει να σημειωθεί ότι ο λόγος για τον οποίο ο χρήστης έχει τη δυνατότητα να πραγματοποιήσει κριτική των αποτελεσμάτων είναι ότι μετά από κάθε αίτηση στο σύστημα το σύνολο συστάσεων που προτείνεται στο χρήστη είναι αρκετά πετυχημένο ώστε να παρουσιάζει τουλάχιστον ένα αντικείμενο το οποίο ενδιαφέρει το χρήστη. Αν το σύστημα δυσλειτουργήσει ή δεν καταφέρει να αντιληφθεί τα στοιχεία που ενδιαφέρουν το χρήστη και δώσει κακές συστάσεις, τότε ο χρήστης δεν μπορεί να προβεί σε κριτική και η διαδικασία παροχής συστάσεων θεωρείται αποτυχημένη. Κατά κανόνα, εάν οι συστάσεις είναι πετυχημένες και ο χρήστης πραγματοποιήσει κριτική των αποτελεσμάτων, τα νέα αποτελέσματα που θα παρουσιαστούν θα είναι λιγότερα από τα προηγούμενα. Παρόλα αυτά, υπάρχει η δυνατότητα σχεδίασης συστημάτων σύστασης βασισμένων σε περιπτώσεις όπου τα αποτελέσματα δε μειώνονται μετά από κάθε επανάληψη, αλλά αντιθέτως παραμένουν στάσιμα ή αυξάνονται προκειμένου να διευρύνουν τα όρια της αναζήτησης. Αυτός ο τύπος σχεδίασης έχει ξεχωριστά πλεονεκτήματα και μειονεκτήματα. Από τη μία πλευρά, ο χρήστης μπορεί να εξετάσει αντικείμενα ευρύτερου ενδιαφέροντος τα οποία δεν αντικατοπτρίζουν επακριβώς τα χαρακτηριστικά του αρχικού παραδείγματός του και επομένως να βρει ένα αντικείμενο που του αρέσει με ιδιαίτερα χαρακτηριστικά, αλλά από την άλλη υπάρχει ο κίνδυνος της παρουσίασης όλων και πιο άσχετων με τα ενδιαφέροντα του χρήστη αντικείμενα.

Μέσω της επαναλαμβανόμενης κριτικής των αποτελεσμάτων, ο χρήστης μπορεί μερικές φορές να φτάσει σε ένα αποτέλεσμα που είναι αρκετά διαφορετικό από το αρχικό του παράδειγμα. Εξάλλου, είναι πολύ δύσκολο για ένα χρήστη να μπορέσει να δώσει ένα ικανοποιητικό παράδειγμα ή να περιγράψει επακριβώς τα χαρακτηριστικά που τον ενδιαφέρουν με την πρώτη προσπάθεια. Κι αυτό γιατί πολλές φορές ο χρήστης μπορεί να έχει συγκεκριμένη αντίληψη για τα αντικείμενα του πεδίου ενδιαφέροντος που επιθυμεί να λάβει ως συστάσεις, αλλά στην πραγματικότητα η αντίληψη αυτή να μην περιγράφει ικανοποιητικά το σύνολο των αντικειμένων που τον ενδιαφέρουν. Παρέχοντας τη δυνατότητα εμπλουτισμού ή επανέναρξης της διαδικασίας των συστάσεων αυξάνεται η πιθανότητα ο χρήστης να ικανοποιηθεί από τη χρήστη της μηχανής συστάσεων. Για παράδειγμα, ένας

χρήστης που χρησιμοποιεί ένα σύστημα συστάσεων ταινιών μπορεί να ξεκινήσει δηλώνοντας ως ταινίες που τον ενδιαφέρουν τις ταινίες Red και Die Hard και να αναμένει να λάβει ως αποτελέσματα άλλες ταινίες δράσεις με πρωταγωνιστή τον Bruce Willis. Αν το σύστημα δουλέψει ικανοποιητικά, θα λάβει πράγματι τέτοιου είδους ταινίες, αλλά θα λάβει και άλλες, όπως για παράδειγμα το Transporter και το Crank. Βλέποντας ο χρήστης τις συγκεκριμένες ταινίες τού δίνεται η δυνατότητα να επαναπροσδιορίσει τις προσδοκίες του για τα προτεινόμενα αποτελέσματα και να επαναλάβει τη διαδικασία των συστάσεων αυτή τη φορά προσπαθώντας να εντοπίσει άλλες ταινίες δράσης με πρωταγωνιστή τον Jason Statham ή/και ταινίες δράσης με πρωταγωνιστές και τον Jason Statham και τον Bruce Willis (πχ το Expendables 3).

Προκειμένου τα συστήματα σύστασης βασισμένα σε περιπτώσεις να δουλέψουν ικανοποιητικά, υπάρχουν δύο σημαντικές πλευρές του συστήματος που θα πρέπει να σχεδιαστούν αποτελεσματικά:

1. Μετρικές ομοιότητας: Η αποτελεσματική σχεδίαση των μετρικών ομοιότητας είναι πολύ σημαντική στα συστήματα συστήματα σύστασης βασισμένα σε περιπτώσεις. Η σημασία των διαφόρων χαρακτηριστικών θα πρέπει να λαμβάνεται υπόψη κατά τη σχεδίαση των μετρικών προκειμένου το σύστημα να λειτουργεί αποτελεσματικά.

2. Μέθοδοι της κριτικής αποτελεσμάτων: Η διαδραστική εξερεύνηση του πεδίου των αντικειμένων ενισχύεται από τις μεθόδους κριτικής. Μία ποικιλία διαφορετικών μεθόδων είναι διαθέσιμη για την κάλυψη διαφορετικών διερευνητικών αναγκών.

Οι μετρικές ομοιότητας θα αναλυθούν σε επόμενο κεφάλαιο αναλυτικά. Παρακάτω θα γίνει σύντομη αναφορά στα βασικά χαρακτηριστικά των μεθόδων κριτικής και σε ορισμένα απλά μοντέλα τους.

1.3.3.4 Μέθοδοι Κριτικής Αποτελεσμάτων

Όπως έχει αναφερθεί, βασικός λόγος για την εισαγωγή των μεθόδων κριτικής αποτελεσμάτων είναι η πεποίθηση ότι ο χρήστης δεν είναι σε θέση να δηλώσει το σύνολο των απαιτήσεών του από το αρχικό ερώτημα. Σε ορισμένα πολύπλοκα πεδία ενδιαφέροντος, μπορεί ακόμα και να είναι δύσκολο να

μεταφραστούν οι ανάγκες του σε χαρακτηριστικά των αντικειμένων με σημασιολογικά ουσιώδη τρόπο. Τις περισσότερες φορές μόνο αφότου ο χρήστης έχει παρατηρήσει τα αποτελέσματα ενός ερωτήματος ο συγκεκριμένος χρήστης αντιλαμβάνεται τον ακριβή τρόπο με τον οποίο μπορεί να δομήσει το ερώτημά του για να λάβει ικανοποιητικά αποτελέσματα. Οι μέθοδοι κριτικής αποτελεσμάτων σχεδιάζονται για να δώσουν στο χρήστη, μεταξύ άλλων, και αυτή τη δυνατότητα.

Αφότου τα αποτελέσματα έχουν παρουσιαστή στους χρήστες, η ανάδραση που λαμβάνει το σύστημα πραγματοποιείται μέσω της κριτικής. Σε πολλές περιπτώσεις, οι διαπροσωπείες έχουν σχεδιαστεί για να διευκολύνουν την εφαρμογή κριτικής στο πιο παρόμοιο στοιχείο ταυτοποίησης, αν και είναι τεχνικά δυνατό για τον χρήστη να επικρίνει κάποιο άλλο από τα στοιχεία της ανακτημένης λίστας k στοιχείων. Σε κριτικές, οι χρήστες καθορίζουν αιτήματα αλλαγής για ένα ή περισσότερα χαρακτηριστικά ενός αντικειμένου που μπορεί να τους αρέσουν. Για παράδειγμα, στην εφαρμογή οικιακής χρήσης του Σχήματος, ο χρήστης μπορεί να προτιμήσει ένα συγκεκριμένο σπίτι, αλλά μπορεί να θέλει το σπίτι σε διαφορετική τοποθεσία ή με ένα ακόμα υπνοδωμάτιο. Ως εκ τούτου, ο χρήστης μπορεί να καθορίσει τις αλλαγές στα χαρακτηριστικά ενός από τα στοιχεία που του αρέσει. Ο χρήστης μπορεί να καθορίσει μια κατευθυντήρια κριτική (π.χ. "φθηνότερη") ή μια κριτική αντικατάστασης (π.χ. "διαφορετικό χρώμα"). Σε αυτές τις περιπτώσεις, εξαλείφονται παραδείγματα που δεν ικανοποιούν τις καθορισμένες από το χρήστη κριτικές και ανακτώνται παραδείγματα παρόμοια με το στοιχείο που προτιμά ο χρήστης (αλλά ικανοποιούν την τρέχουσα ακολουθία κριτικών). Όταν καθορίζονται πολλαπλές κριτικές σε διαδοχικούς κύκλους συστάσεων, προτιμώνται πιο πρόσφατες κριτικές.

Σε δεδομένη χρονική στιγμή, ο χρήστης μπορεί να καθορίσει είτε ένα μόνο χαρακτηριστικό είτε ένα συνδυασμό χαρακτηριστικών για τροποποίηση. Στο πλαίσιο αυτό, οι κριτικές είναι τριών διαφορετικών τύπων, που αντιστοιχούν σε απλές κριτικές, σύνθετες κριτικές, ολικές κριτικές και δυναμικές κριτικές:

1. Στις απλές κριτικές ο χρήστης προσδιορίζει μια μόνο αλλαγή σε ένα από τα χαρακτηριστικά ενός προτεινόμενου αντικειμένου.

2. Στις σύνθετες κριτικές ο χρήστης μπορεί να προσδιορίσει όσες αλλαγές θέλει σε όσα χαρακτηριστικά θέλει σε κάποιο προτεινόμενο αντικείμενο.

3. Στις ολικές κριτικές όλα τα χαρακτηριστικά του αντικειμένου με βάση το οποίο έγινε η αρχική αναζήτηση αντικαθίστανται από τα χαρακτηριστικά του νέου επιλεγμένου αντικειμένου.

4. Στις δυναμικές κριτικές ο στόχος είναι να χρησιμοποιηθούν μέθοδοι εξόρυξης δεδομένων για τα ανακτηθέντα αποτελέσματα ώστε να προσδιοριστουν οι πιο καρποφόρες κατευθύνσεις της εξερεύνησης και να παρουσιαστούν στο χρήστη. Έτσι, οι δυναμικές κριτικές είναι, εξ ορισμού, σύνθετες κριτικές επειδή σχεδόν πάντα αντιπροσωπεύουν συνδυασμούς αλλαγών που παρουσιάζονται στο χρήστη. Η κύρια διαφορά είναι ότι παρουσιάζεται μόνο το υποσύνολο των πιο σχετικών δυνατοτήτων, με βάση τα επί του παρόντος ανακτηθέντα αποτελέσματα. Ως εκ τούτου, οι δυναμικές κριτικές έχουν σχεδιαστεί για να παρέχουν καλύτερη καθοδήγηση στο χρήστη κατά τη διάρκεια της διαδικασίας αναζήτησης.

2 Οντολογική Αναπαράσταση Γνώσης και Συλλογιστική

Περίληψη. Σε αυτό το κεφάλαιο αναλύονται βασικές έννοιες που σχετίζονται με την οντολογική αναπαράσταση γνώσης με ιδιαίτερη έμφαση στον τρόπο με τον οποίο εμφανίζεται η γνώση στον Παγκόσμιο Ιστό. Εισάγονται στοιχεία όπως οντολογία, βάση γνώσης, έννοια, άτομο, αντικείμενο, ρόλοι και άλλα. Επίσης, αναφέρονται οι βασικοί κατασκευαστές μιας οντολογίας και αναλύεται η διαδικασία απάντησης συζευτικών ερωτημάτων.

Σε πολλές περιπτώσεις είναι χρήσιμη η αναπαράσταση της γνώσης με τη μορφή κατηγοριών αντικειμένων. Ξεκινώντας από τον καθορισμό των αντικειμένων, μέσω της απόδοσης ονομάτων, είναι χρήσιμο να μελετήσουμε τις ιδιότητές τους και να περιγράψουμε τον τρόπο με τον οποίο σχετίζονται μεταξύ τους και καθορίζουν τα χαρακτηριστικά των αντικειμένων. Με τον τρόπο αυτό, είναι χρήσιμο η αναπαράσταση της γνώσης του πεδίου ενδιαφέροντος να κωδικοποιεί και να ενσωματώνει τις παρακάτω (διαισθητικές) σημασιολογικές παρατηρήσεις[2]:

- Τα αντικείμενα έχουν συγκεκριμένα χαρακτηριστικά και με βάση αυτά κατατάσσονται σε διάφορες κατηγορίες. Για παράδειγμα μία ταινία μπορεί να είναι Action, Adventure, Animation ή Comedy.
- Ορισμένα αντικείμενα μπορεί να ανήκουν σε κατηγορίες οι οποίες είναι υποκατηγορίες άλλων κατηγοριών. Για παράδειγμα η κατηγορία Super-Hero Movie είνα ι υποκατηγορία της κατηγορίας Action Movie. Επομένως, όλες οι ταινίες που είναι Super-Hero Movies είναι επίσης και Action Movies.
- Συνήθως οι κατηγορίες έχουν ονόματα, δηλαδή απλές αναφορές (Comedy, Actor, Movie). Σε πολλές, όμως, από τις κατηγορίες αναφερόμαστε περιγραφικά, με σύνθετα ονόματα που συνήθως περιγράφουν λεκτικά τις ιδιότητές που πρέπει να έχουν τα αντικείμενα

που κατατάσσονται στη συγκεκριμένη κατηγορία (για παράδειγμα “ταινίες με πρωταγωνιστή τον Bruce Willis”).

- Οι σχέσεις ενός αντικειμένου με άλλα αντικείμενα συγκεκριμένης κατηγορίας είναι σημαντικές για την κατάταξή τους σε κάποια κατηγορία (π.χ. οι ταινίες που έχουν μεγάλη διάρκεια ονομάζονται μεγάλου μήκους) .
- Κάποια αντικείμενα μπορεί να αποτελούνται από άλλα (π.χ. κάθε πλάνο μιας ταινίας είναι μέρος μιας σκηνής).
- Οι σχέσεις των μερών ενός αντικειμένου είναι πιθανώς σημαντικές για την κατάταξή του σε μια κατηγορία.

Περιγραφές που διατυπώνουν πληροφορίες αυτού του τύπου για ένα συγκεκριμένο πεδίο ονομάζονται συνήθως ορολογικές, ενώ οι αντίστοιχες γνώσεις που αναπτύσσονται ονομάζονται ορολογίες ή οντολογίες. Η σημασιολογική ερμηνεία και η λογική ανάλυση των οντολογιών, με βάση τις παραπάνω παρατηρήσεις, χρησιμοποιείται συνήθως για να μπορέσουμε να προσεγγίσουμε προβλήματα όπως το αν κάποια κατηγορία είναι γενικότερη από κάποια άλλη (πχ αν οι Super-Hero Movies είναι Action Movies), αν κάποιο συγκεκριμένο αντικείμενο ανήκει σε μια κατηγορία (πχ αν σε μία ταινία πρωταγωνιστεί κάποιος ηθοποιός) και άλλα. Δεν είναι δύσκολο να δούμε ότι η μοντελοποίηση του πεδίου ενδιαφέροντος που υιοθετείται κατά την οντολογική αναπαράσταση γνώσης είναι αντικειμενοστρεφής. Συγκεκριμένα, τα άτομα (απτά ή αφηρημένα αντικείμενα) αποτελούν τα βασικά στοιχεία του κόσμου. Οι ιδιότητές τους περιγράφονται με τη βοήθεια των κατηγοριών και των ταξινομήσεών τους σε αυτές, αλλά και των συσχετίσεων μεταξύ τους.

Στη βιβλιογραφία έχουν προταθεί και αναπτυχθεί αρκετοί αντικειμενοστρεφείς φορμαλισμοί αναπαράστασης γνώσης (όπως τα σημασιολογικά δίκτυα (semantic networks), τα πλαίσια (frames) κλπ). Οι περισσότεροι, όμως, από αυτούς δεν ενσωματώνουν στοιχεία της μαθηματικής λογικής για την τυπική μελέτη και ανάλυση των περιορισμών που διατυπώνονται. Συνεπώς, περιορίζονται στην τυπική αναπαράσταση της μοντελοποίησης του κόσμου, όχι απαραίτητα με στόχο την ανάπτυξη συστημάτων αυτόματης συλλογιστικής (για παράδειγμα, τόσο τα σημασιολογικά δίκτυα όσο και τα πλαίσια χρησιμοποιούνται ως φορμαλισμοί

μοντελοποίησης συστημάτων λογισμικού κατά τη διαδικασία ανάπτυξης λογισμικού). Τα τελευταία χρόνια, ειδικά με την ανάπτυξη του Παγκόσμιου Ιστού, έχει καταγραφεί η ανάγκη για την ανάπτυξη φορμαλισμών αναπαράστασης, διαχείρισης και χρήσης οντολο- γικής γνώσης που να υποστηρίζονται από διαδικασίες αυτόματης συλλογιστικής οι οποίες να μπορούν να εφαρμοστούν σε πρακτικά προβλήματα. Στο πλαίσιο αυτό, έχουν προταθεί και μελετηθεί οι περιγραφικές λογικές (Description Logics), ένας τυπικός φορμαλισμός αναπαράστασης γνώσης με βάση τον οποίο οι μηχανικοί γνώσης αναπτύσσουν ορολογίες που υποστηρίζονται από αλγόριθμους συλλογιστικής με ικανοποιητικά υπολογιστικά χαρακτηριστικά.

2.1 Τυπικές ορολογικές περιγραφές

Το πρώτο βήμα κατά την ανάπτυξη τυπικών ορολογικών περιγραφών είναι η ταυτοποίηση και θεμελίωση της αναφοράς στα *άτομα* (individuals – αναφερόμενα και ως *αντικείμενα*), τα στοιχεία, δηλαδή, του κόσμου που περιγράφουμε. Για το σκοπό αυτό, αντιστοιχίζουμε τα άτομα σε συμβολοσειρές (ονόματα), οι οποίες αποτελούν συντακτικούς προσδιοριστές που δίνουν τη δυνατότητα τυπικής, μονοσήμαντης αναφοράς σε αυτά. Για να είναι, όμως, η αναφορά μονοσήμαντη (στοιχείο απαραίτητο στις τυπικές περιγραφές), κάθε όνομα πρέπει να αντιστοιχεί σε ένα μόνο άτομο. Αντιθέτως, τα άτομα δεν είναι απαραίτητο να έχουν μοναδικά ονόματα. Η δέσμευση για την ισχύ του συγκεκριμένου περιορισμού ονομάζεται *υπόθεση μοναδικού ονόματος* (unique name assumption, UNA). Η υπόθεση αυτή, αν και απλοποιεί αρκετά τη διαδικασία αναφοράς, σε πολλές περιπτώσεις είναι περιοριστική. Δεν είναι ρεαλιστικό, για παράδειγμα, να υποθέσουμε ότι στον Παγκόσμιο Ιστό θα υπάρχει μοναδικό όνομα (μοναδικός προσδιοριστής) για την Αθήνα, ο οποίος θα χρησιμοποιείται σε όλες τις αναφορές σε αυτή (για παράδειγμα σε όλες τις ιστοσελίδες του Παγκόσμιου Ιστού), ή ότι ο τίτλος “A Pure Formality” και ο τίτλος “Una Pura Formalita” (ο ιταλικός τίτλος της ταινίας με πρωταγωνιστή τον Gerard Depardieu) αναφέρονται, υποχρεωτικά, σε διαφορετικά αντικείμενα του κόσμου. Από την άλλη πλευρά, αν δεχθούμε ότι δεν ισχύει η υπόθεση μοναδικού ονόματος, είναι πολύ χρήσιμο να μπορούμε να δηλώσουμε ότι δύο ονόματα αναφέρονται στο ίδιο άτομο. Τυπικά, αυτό θα μπορούσε δηλωθεί με την έκφραση:

$$A\,Pure\,Formality \approx Una\,Pura\,Formalita \quad (1.1)$$

η οποία ονομάζεται ισότητα ατόμων (individual equality).

Αντίστοιχα, θα ήταν χρήσιμο να μπορούμε να δηλώσουμε ότι οι ταινίες στις οποίες αναφέρονται οι τίτλοι “A Pure Formality” και “Die Hard” είναι διαφορετικές (αν και είναι διαισθητικά προφανές). Τυπικά, χρησιμοποιούμε την έκφραση:

$$A\,Pure\,Formality \neq Die\,Hard \quad (1.2)$$

η οποία ονομάζεται ανισότητα ατόμων (individual inequality).

Στη συνέχεια, θα δούμε με ποιον τρόπο μπορούμε να περιγράφουμε τις ιδιότητες των ονοματισμένων ατόμων. Γενικά, τα άτομα ταξινομούνται σε *κατηγορίες* (categories) ή *κλάσεις* (classes), με βάση τις ιδιότητές τους, όπως ακριβώς τα αντικείμενα ανήκουν ή όχι σε ένα σύνολο, ανάλογα με το αν πληρούν μια συγκεκριμένη ιδιότητα. Για παράδειγμα, μπορούμε να δηλώσουμε τυπικά ότι η ταινία “Die Hard” είναι ταινία δράσης μέσω της έκφρασης:

$$ActionMovie(DieHard) \quad (1.3)$$

στην οποία το όνομα “ActionMovie” χρησιμοποιείται για τον καθορισμό της *έννοιας* (concept) της ταινίας δράσης. Αυτό σημαίνει ότι η ταινία “Die Hard” έχει όλα τα χαρακτηριστικά της έννοιας “ActionMovie”. Στην περίπτωση αυτή λέμε ότι η ταινία “Die Hard” είναι στιγμιότυπο (instance) της έννοιας “ActionMovie”.

Με παρόμοιο τρόπο αναπαριστούμε τις σχέσεις που έχουν τα άτομα μεταξύ τους. Για παράδειγμα, δηλώνουμε τυπικά ότι το άτομο Bruce Willis είναι πρωταγωνιστής της ταινίας “Die Hard” μέσω της έκφρασης:

$$isProtagonist(BruceWillis, DieHard) \quad (1.4)$$

στην οποία το όνομα “isProtagonist” χρησιμοποιείται για τον καθορισμό της *σχέσης* (relation) “είναι-πρωταγωνιστής-του”, όπου το πρώτο όρισμα είναι πρωταγωνιστής του δεύτερου ορίσματος. Στις περιγραφικές λογικές (που αποτελούν τυπικό αντικειμενοστρεφές μοντέλο αναπαράστασης γνώσης) οι σχέσεις χρησιμοποιούνται για να περιγράψουν ιδιότητες των ατόμων που εμπλέκουν άλλα άτομα και είναι πάντα δυαδικές (έχουν δύο ορίσματα). Για αυτόν τον λόγο αναφέρονται και ως *ρόλοι* (roles).

Η περιγραφή των ατόμων μέσω δηλώσεων των παραπάνω μορφών είναι παρόμοια με την αναπαράσταση που χρησιμοποιείται στις βάσεις δεδομένων. Μία προφανής διαφορά είναι ότι στην περίπτωση των περιγραφικών λογικών δηλώνεται ρητά (explicitly) και η σχέση των ονομάτων. Από την άλλη πλευρά, στα σχεσιακά μοντέλα βάσεων δεδομένων μπορούν να αναπαρασταθούν σχέσεις με περισσότερα από δύο ορίσματα. Η σημαντικότερη όμως διαφορά τους προκύπτει από τη σημασιολογική ερμηνεία των δύο τελευταίων δηλώσεων, η οποία θα περιγραφεί παρακάτω.

Οι κατηγορίες στις οποίες ταξινομούνται τα άτομα ονοματίζονται με τη χρήση συμβολοσειρών (όπως είδαμε και προηγουμένως) και αντιστοιχούν σε μια συγκεκριμένη ιδιότητα, την οποία εξετάζουμε για να δούμε αν κάποιο άτομο είναι μέλος αυτής της κατηγορίας ή όχι. Η ιδιότητα αυτή μπορεί να είναι, διαισθητικά, από απλή έως και ιδιαίτερα περίπλοκη, ανάλογα με την έννοια που αντιστοιχεί στη συγκεκριμένη κατηγορία.

Επιπλέον, κάποιες από τις ιδιότητες αυτές μπορεί να σχετίζονται. Για παράδειγμα η ιδιότητα που καθορίζει την κατηγορία των ταινιών δράσης και την αντίστοιχη έννοια “ActionMovie” είναι γενικότερη από την ιδιότητα που καθορίζει την κατηγορία των ταινιών με υπερήρωες, αφού όλες οι ταινίες με υπερήρωες είναι και ταινίες δράσης. Σε αυτή την περίπτωση, θα λέμε ότι η έννοια “SuperHeroMovie” είναι *υποέννοια* της έννοιας “ActionMovie”. Τυπικά, γράφουμε:

$$SuperHeroMovie \sqsubseteq ActionMovie \quad (1.5)$$

υπονοώντας ότι κάθε άτομο το οποίο είναι “SuperHeroMovie” είναι και ActionMovie. Στην περίπτωση αυτή λέμε ότι η έννοια “SuperHeroMovie” *υπάγεται* (subsumed) στην έννοια “ActionMovie”.

Μπορούμε, επίσης, να δηλώσουμε ότι δύο έννοιες είναι ξένες μεταξύ τους, αν οι ιδιότητές τους καθορίζουν κατηγορίες ατόμων που είναι ξένες μεταξύ τους. Για παράδειγμα, οι έννοιες “Movie” και “Actor” είναι ξένες μεταξύ τους, γιατί κανένα άτομο του κόσμου δεν μπορεί να είναι ταυτόχρονα ηθοποιός και ταινία. Τυπικά, δηλώνουμε ότι:

$$Actor \neq Movie \quad (1.6)$$

και λέμε ότι οι έννοιες “ηθοποιός” και “ταινία” είναι *ξένες* (disjoint).

Δύο έννοιες είναι δυνατόν να ταυτίζονται, αν οι ιδιότητές τους καθορίζουν πάντα το ίδιο σύνολο ατόμων. Για παράδειγμα, οι έννοιες “Film” και “Movie“ ταυτίζονται στο πεδίο του κινηματογράφου, διότι τα στιγμιότυπα των φιλμ είναι και ταινίες και αντίστροφα (οι λέξεις φιλμ και ταινία είναι συνώνυμες, διότι ιστορικά το υλικό καταγραφής της κινούμενης εικόνας στον κινηματογράφο ήταν το φιλμ παρότι τα τελευταία χρόνια με τη χρήση των ψηφιακών μέσων αυτό δεν ισχύει πλέον, η ισοδυναμία αυτή έχει επικρατήσει). Η παραπάνω δήλωση γράφεται τυπικά ως εξής:

$$Movie \equiv Film \quad (1.7)$$

Η ισοδυναμία εννοιών δεν περιορίζεται μόνο στα συνώνυμα. Για παράδειγμα, οι έννοιες “Movie” και “ShortFilmOrFeatureFilm” επίσης ταυτίζονται, γιατί διαισθητικά τα στιγμιότυπα των ταινιών είναι ταινίες μικρού μήκους ή ταινίες μεγάλου μήκους και αντίστροφα. Τυπικά, δηλώνουμε ότι:

$$Movie \equiv ShortFilmOrFeatureFilm \quad (1.8)$$

και λέμε ότι οι έννοιες “Movie” και “ShortFilmOrFeatureFilm” είναι ισοδύναμες (equivalent).

Στη συγκεκριμένη σχέση παρατηρούμε ότι η έννοια “ShortFilmOrFeatureFilm” έχει ένα περίπλοκο όνομα, το οποίο ουσιαστικά υποδηλώνει ότι τα στιγμιότυπα της έννοιας είναι ταινίες μικρού μήκους ή ταινίες μεγάλου μήκους. Παρατηρούμε ότι, ενώ τα επίθετα που χρησιμοποιούνται στα ονόματα («μικρό» και «μεγάλο») χαρακτηρίζουν τη συγκεκριμένη περίπτωση, περιγράφοντας τις συγκεκριμένες ιδιότητες, το «ή» είναι ένας σύνδεσμος που χρησιμοποιείται με τον ίδιο τρόπο σε πολλά ονόματα· είναι ένας σύνδεσμος που η σημασιολογία του είναι *λογική*. Ένας τρόπος για να εμπλουτιστεί η γλώσσα αναπαράστασης είναι η χρήση τυπικών (λογικών) συνδέσμων για την περιγραφή της σχέσης που κρύβεται στη συμβολοσειρά του ονόματος.

Για να έχουμε μία καλύτερη διαισθητική περιγραφή μπορούμε να δηλώσουμε:

$$Movie \equiv ShortFilm \sqcup FeatureFilm \quad (1.9)$$

όπου το σύμβολο $\sqcup$ χρησιμοποιείται για να δηλώσει τη *διάζευξη* (disjunction) των εννοιών “ShortFilm” και “FeatureFilm”.

Σύμβολα αυτού του τύπου ονομάζονται *κατασκευαστές* (constructors) και συνδέουν τις ιδιότητες των κατηγοριών των ατόμων που περιγράφονται από τις επιμέρους έννοιες. Παρακάτω παρουσιάζονται οι βασικοί κατασκευαστές που χρησιμοποιούνται στην ανάπτυξη ορολογιών.

Η *σύζευξη* (conjunction) δύο εννοιών προσδιορίζει την κατηγορία των ατόμων που έχουν τις ιδιότητες και των δύο εννοιών. Για παράδειγμα, μια ταινία μεγάλου μήκους αποτελεί ουσιαστικά μια ταινία με μεγάλη διάρκεια. Τυπικά, λέμε ότι η έννοια FeatureFilm είναι η σύζευξη της έννοιας Film και της έννοιας LongFilm (στο πεδίο του κινηματογράφου, ένα φιλμ λέγεται μακρύ, αν είναι μακρύτερο από 1600 μέτρα, δηλαδή αν η ταινία που αποθηκεύει είναι διάρκειας μεγαλύτερης από 40 λεπτά) και γράφουμε:

$$FeatureFilm \equiv Film \sqcap LongFilm \quad (1.10)$$

όπου ⊓ είναι το σύμβολο της σύζευξης.

Η *άρνηση* (negation) μιας έννοιας προσδιορίζει την κατηγορία των ατόμων που δεν έχουν την ιδιότητα της έννοιας. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε την άρνηση για να δηλώσουμε ότι μια ταινία που δεν είναι μικρού μήκους είναι μεγάλου μήκους. Τυπικά, λέμε ότι η έννοια FeatureFilm είναι η σύζευξη της έννοιας Film και της άρνησης της έννοιας ShortFilm και γράφουμε:

$$FeatureFilm \equiv Film \sqcap \neg ShortFilm \quad (1.11)$$

όπου ¬ είναι το σύμβολο της σύζευξης.

Είναι σημαντικό να τονίσουμε ότι, αν ο ορισμός της έννοιας FeatureFilm δινόταν από τη σχέση:

$$FeatureFilm \equiv \neg ShortFilm \quad (1.12)$$

τότε ουσιαστικά θα δηλώναμε ότι κάθε άτομο του κόσμου που δεν είναι στιγμιότυπο της έννοιας ShortFilm (ακόμα και ο Bruce Willis για παράδειγμα) είναι στιγμιότυπο της έννοιας FeatureFilm. Η σύζευξη με την έννοια Film όπως αυτή παρουσιάζεται παραπάνω αποκλείει την παρερμηνεία αυτή.

Σε πολλές περιπτώσεις είναι χρήσιμο να ορίζουμε κατηγορίες ατόμων από τις σχέσεις τους με άλλα άτομα. Για παράδειγμα, μπορούμε να δηλώσουμε ότι ένα

άτομο είναι στιγμιότυπο της έννοιας πρωταγωνιστής αν έχει πρωταγωνιστήσει σε τουλάχιστον μία ταινία. Τυπικά, ορίζουμε την έννοια MovieProtagonist από την ύπαρξη στιγμιοτύπου του ρόλου isProtagonist ως εξής:

$$MovieProtagonist \equiv \exists hasProtagonist.Movie \quad (1.13)$$

όπου ∃ είναι το σύμβολο του *υπαρξιακού ποσοδείκτη* (existential quantifier). Στην περίπτωση αυτή χρησιμοποιούμε την έννοια Movie για να προσδιορίσουμε με ακρίβεια ότι θα λέμε πρωταγωνιστή μόνο όποιον έχει πρωταγωνιστήσει σε κάποια *ταινία* (αποφεύγοντας έτσι πιθανά λάθη στον ορισμό στιγμιοτύπων της σχέσης isProtagonist).

Παρόμοια χρησιμοποιείται ο *καθολικός ποσοδείκτης* (universal quantifier) για να δηλώσουμε ότι κάποια άτομα σχετίζονται μόνο με άτομα μιας συγκεκριμένης κατηγορίας, μέσω μιας συγκεκριμένης σχέσης. Για παράδειγμα, δηλώνουμε ότι στις ταινίες πρωταγωνιστούν ηθοποιοί. Τυπικά, γράφουμε:

$$Movie \sqsubseteq \forall \ hasProtagonist.Actor \quad (1.14)$$

εννοώντας ότι αν τα στιγμιότυπα της έννοιας Movie συνδέονται με κάποιο άτομο μέσω της σχέσης hasProtagonist, το άτομο αυτό θα είναι στιγμιότυπο της έννοιας Actor.

Σε ορισμένες περιπτώσεις είναι χρήσιμο να προσδιορίσουμε με μεγαλύτερη ακρίβεια τον αριθμό των ατόμων με τα οποία σχετίζονται τα άτομα. Για παράδειγμα, μπορούμε να ορίσουμε τους δημοφιλείς ηθοποιούς ως τους ηθοποιούς εκείνους που έχουν πρωταγωνιστήσει σε πάνω από 20 ταινίες. Τυπικά, δηλώνουμε ότι:

$$PopularActor \equiv \geq 20 \text{isProtagonist.Movie} \quad (1.15)$$

Στην περίπτωση αυτή θα λέμε ότι έχουμε έναν *περιορισμό ελάχιστης πληθικότητας με εξειδίκευση* (qualified number restriction). Αντίστοιχα ορίζεται και ο περιορισμός μέγιστης πληθικότητας με εξειδίκευση.

Στη δήλωση (1.14) παρατηρούμε ότι περιορίζουμε την ιδιότητα της σχέσης hasProtagonist να λαμβάνει τιμές από την κατηγορία που ορίζει η έννοια Actor μόνο για τα στιγμιότυπα της έννοιας Movie, ενώ κάτι τέτοιο ισχύει για όλα τα

άτομα του κόσμου που ανήκουν στο πεδίο της hasProtagonist. Για να το ορίσουμε αυτό, θα έπρεπε να αναφερθούμε σε όλα τα άτομα του κόσμου, δηλαδή να ορίσουμε μια έννοια που έχει στιγμιότυπα όλα τα άτομα του κόσμου, σε όλες τις ορολογίες. Η έννοια αυτή συμβολίζεται με τον τελεστή $\top$ (την ονομάζουμε Top). Στο προηγούμενο παράδειγμα θα δηλώναμε ότι:

$$\top \sqsubseteq \forall \ hasProtagonist.Actor \quad (1.16)$$

Με αντίστοιχο τρόπο ορίζεται η έννοια που δεν μπορεί να έχει στιγμιότυπα σε κανέναν κόσμο, δηλαδή κανένα άτομο δεν μπορεί να ικανοποιεί την ιδιότητα ορισμού της κατηγορίας που ορίζουν. Τη συμβολίζουμε με τον τελεστή $\bot$ και την ονομάζουμε Bottom. Η έννοια αυτή θα μπορούσε για παράδειγμα να χρησιμοποιηθεί σε δηλώσεις του τύπου:

$$FeatureFilm \sqcap ShortFilm \sqsubseteq \bot \quad (1.17)$$

με την οποία δηλώνεται ότι αν ένα άτομο είναι στιγμιότυπο της έννοιας FeatureFilm και, ταυτόχρονα, στιγμιότυπο της έννοιας ShortFilm, τότε είναι και στιγμιότυπο της έννοιας Bottom, γεγονός αδύνατο.

Συνεπώς οι έννοιες FeatureFilm και ShortFilm δεν μπορούν να έχουν κοινό στιγμιότυπο, άρα οι κατηγορίες ατόμων που ορίζουν είναι ξένες μεταξύ τους. Η δήλωση αυτή θα μπορούσε να γίνει και ως εξής:

$$FeatureFilm \neq ShortFilm$$

Ολοκληρώνοντας τη μελέτη των κατασκευαστών εννοιών, θα δούμε μια ειδική κατηγορία ατόμων που ορίζεται από τα ίδια τα μέλη της. Έστω, για παράδειγμα, ότι θέλουμε να περιγράψουμε την προέλευση των ταινιών. Πιο συγκεκριμένα, ας ξεκινήσουμε από την ήπειρο προέλευσής τους. Είναι προφανές ότι τα μοναδικά στιγμιότυπα της έννοιας Continent είναι τα asia, africa, america, antarctica, europe και australia. Για να διασφαλίσουμε τον περιορισμό αυτό, μπορούμε να δηλώσουμε ότι

$$Continent \equiv asia \sqcup \cdots \sqcup australia$$

συμπεριλαμβάνοντας και τις έξι ηπείρους. Στην περίπτωση αυτή, με {asia} συμβολίζουμε την έννοια που έχει μόνο ένα στιγμιότυπο, το άτομο asia (αντίστοιχα και για τις υπόλοιπες ηπείρους). Η έννοια αυτή ονομάζεται *ονοματική* (nominal).

Πολλές φορές είναι χρήσιμο να μπορούμε να δηλώσουμε περιορισμούς που αφορούν τις σχέσεις. Για παράδειγμα, παρατηρούμε ότι όποτε ένα άτομο σχετίζεται με κάποιο άλλο μέσω της σχέσης “hasProtagonist”, σχετίζεται με το ίδιο άτομο μέσω της “hasActor”. Μπορούμε να δηλώσουμε τυπικά τον συγκεκριμένο περιορισμό, μέσω της εξής έκφρασης:

$$hasProtagonist \sqsubseteq hasActor \quad (1.18)$$

Σε αυτήν την περίπτωση λέμε ότι ο ρόλος “hasProtagonist” είναι *υπορόλος* (subrole) του ρόλου “hasActor”.

Επιπλέον, είναι χρήσιμο να εκφράζονται δηλώσεις στις οποίες αντιστρέφονται οι ρόλοι μιας σχέσης. Για παράδειγμα, όταν ένας ηθοποιός a *έχει πρωταγωνιστήσει* σε μια ταινία b, τότε η ταινία b *έχει πρωταγωνιστή* τον a. Η δήλωση αυτή μπορεί να γίνει τυπικά μέσω της έκφρασης:

$$isProtagonist \sqsubseteq hasProtagonist^{-} \quad (1.19)$$

Χρησιμοποιούμε το σύμβολο $(\cdot)^{-}$ για να δηλώσουμε τη συσχέτιση των ρόλων hasProtagonist και isProtagonist, και λέμε ότι οι ρόλοι αυτοί είναι *ανάστροφοι* (inverse). Προφανώς, σε πολλές περιπτώσεις δεν είναι προφανές από τα ονόματα των ρόλων ότι είναι ανάστροφοι (όπως για παράδειγμα οι ρόλοι “hasParent” και “hasChild”).

Πιο περίπλοκες συσχετίσεις ορίζονται μεταξύ περισσότερων από δύο ρόλους. Για παράδειγμα, μπορούμε να δηλώσουμε ότι αν κάποιος σκηνοθέτης έχει σκηνοθετήσει μια ταινία στην οποία έχει παίξει κάποιος ηθοποιός, τότε ο σκηνοθέτης έχει συνεργαστεί με τον συγκεκριμένο ηθοποιό. Τυπικά:

$$isDirector \circ hasActor \sqsubseteq hasCollaboration \quad (1.20)$$

όπου με το σύμβολο $\circ$ εκφράζουμε τη *σύνθεση ρόλων* (role composition).

Τέλος, είναι σημαντικό να μπορούμε να δηλώσουμε τις ιδιότητες που έχουν πολλές φορές οι σχέσεις. Για παράδειγμα, η σχέση hasPart είναι μεταβατική: αν μια ταινία έχει μέρος της μια σκηνή, που έχει μέρος της ένα πλάνο, τότε η ταινία έχει μέρος της το πλάνο. Η δήλωση αυτή γίνεται τυπικά ως εξής:

$$trans(hasPart) \quad (1.21)$$

όπου με το σύμβολο trans συμβολίζουμε τη *μεταβατικότητα* (transitivity). Παρόμοια ορίζονται και άλλες ιδιότητες, όπως η *συμμετρικότητα* (symmetry) (μέσω της οποίας μπορούμε να δηλώσουμε ότι τα άτομα σχετίζονται με μία συγκεκριμένη σχέση σε ζεύγη χωρίς να έχει σημασία η φορά), η *ανακλαστικότητα* (reflexivity) (μέσω της οποίας μπορούμε να δηλώσουμε ότι κάθε άτομο σχετίζεται με τον εαυτό του με τη συγκεκριμένη σχέση) κλπ.

Έχοντας υπόψη μας όλα τα παραπάνω, ένα σύνολο από ισχυρισμούς ισότητας ή ανισότητας στιγμιοτύπων, ισχυρισμούς εννοιών και ισχυρισμούς ρόλων ονομάζεται σώμα ισχυρισμών (assertion box, ABox). Το σύνολο των ονομάτων που χρησιμοποιούνται στους ισχυρισμούς ενός Abox A ονομάζεται υπογραφή του ABox και συμβολίζεται με Sig(A).

Για παράδειγμα, ένα Abox θα μπορούσε να είναι το εξής:

$$\alpha_1.\ Movie(Die\ Hard)$$
$$a_2.\ Movie(Die\ Hard\ 2)$$
$$a_3.\ Movie(Titanic)$$
$$a_4.\ Action(Die\ Hard)$$
$$a_5.\ Action(Die\ Hard\ 2)$$
$$a_6.\ Drama(Titanic)$$
$$a_7.\ Drama(The\ Story\ of\ Us)$$
$$a_8.\ hasProtagonist(The\ Story\ of\ Us, Bruce\ Willis)$$
$$a_9.\ hasProtagonist(Die\ Hard, Bruce\ Willis)$$
$$a_{10}.\ hasProtagonist(Die\ Hard\ 2, Bruce\ Willis)$$
$$a_{11}.\ hasProtagonist(Titanic, Kate\ Winslet)$$

το οποίο έχει ως υπογραφή:

Sig(A) = [Die Hard, Die Hard 2, Titanic, The Story of Us, Bruce Willis, Kate Winslet, hasProtagonist, Action, Drama, Movie]

Σώμα ορολογίας (terminological box, TBox) ή οντολογία (ontology) ονομάζεται ένα σύνολο από αξιώματα υπαγωγής και ισοδυναμίας εννοιών και ρόλων. Το σύνολο των ονομάτων που χρησιμοποιούνται στα αξιώματα ενός Tbox T ονομάζεται υπογραφή του T και συμβολίζεται με Sig (T).

Για παράδειγμα, ένα Tbox θα μπορούσε να είναι το εξής:

$$\tau_1.\ Protagonist \sqsubseteq Actor$$
$$\tau_2.\ Movie \equiv Film$$
$$\tau_3.\ Actor \sqcap Movie \sqsubseteq \bot$$
$$\tau_4.\ Actor \equiv isActor.Movie$$
$$\tau_5.\ Protagonist \equiv isProtagonist.Movie$$
$$\tau_6.\ PopularActor \equiv \geq 20 \text{isActor.Movie}$$

το οποίο έχει ως υπογραφή:

Sig(T) = [Protagonist, Actor, Movie, Film, isActor, isProtagonist, Popular Actor]

Έστω IN, CN και RN ένα σύνολο ονομάτων ατόμων, εννοιών και ρόλων, αντίστοιχα. Θα ονομάζουμε *βάση γνώσης* ή απλά *γνώση K* μια δυάδα *K = (T, A)*, όπου T ένα σώμα ορολογίας και A ένα σώμα ισχυρισμών, με Sig(T), Sig(A)⊆IN ∪ CN ∪ RN. Θα ονομάζουμε *υπογραφή*, Sig(K), της K το σύνολο Sig(T)∪Sig(A).

Από όλα τα παραπάνω είναι προφανές ότι κατά τον ορισμό μιας βάσης γνώσης μπορούν να χρησιμοποιηθούν διάφοροι κατασκευαστές σύνθετων εννοιών και ρόλων στα αξιώματα του σώματος ορολογίας. Μπορούμε επομένως ανάλογα με την εκφραστικότητα που θέλουμε να προσδώσουμε στη βάση γνώσης μας να επιλέξουμε διαφορετικό σύνολο κατασκευαστών για να χρησιμοποιήσουμε. Όσο περισσότερους κατασκευαστές χρησιμοποιήσουμε, τόσο πιο εκφραστική γίνεται η γλώσσα που χρησιμοποιούμε, αλλά ταυτόχρονα, τόσο πιο πολύπλοκη γίνεται η σύνταξη της γνώσης, δυσχεραίνοντας έτσι το έργο της αυτόματης συλλογιστικής.

Με βάση το σύνολο των κατασκευαστών που χρησιμοποιούνται, οι γλώσσες που προκύπτουν έχουν διαφορετικά ονόματα. Οι πιο σημαντικές από αυτές τις γλώσσες είναι οι παρακάτω ALC, AL, SI, SHI, SHIN, SHOIN, SHOIQ, SROIQ.

2.2 Απλή Συλλογιστική σε Συζευτικά Ερωτήματα[3]

Μέχρι τώρα έχει περιγραφεί ο τρόπος με τον οποίο ορίζεται και συντάσσεται μία βάση γνώσης. Από μόνη της ωστόσο, η συγκεκριμένη αναπαράσταση του πεδίου ενδιαφέροντος, παρότι σημασιολογικά ορθή και ακριβής, λίγη αξία έχει σε πρακτικές εφαρμογές. Αυτό το οποίο έχει σημασία είναι η δυνατότητα αξιοποίησης της βάσης για τη λήψη απαντήσεων σε ερωτήματα των οποίων η

απάντηση ενδεχομένως να μην είναι άμεσα εμφανής από απλή παρατήρηση των διαθέσιμων δεδομένων.

Έστω K(T, A), μια βάση γνώσης, όπου T ένα σώμα ορολογίας και A ένα σώμα υποθέσεων, με IN, CN και RN το σύνολο των ονομάτων των ατόμων, εννοιών και ρόλων της K αντίστοιχα. Θα καλούμε ατομικό ερώτημα (atomic query) μια έκφραση της μορφής $q_1(x) = C(x)$ ή της μορφής $q_2(x, y) = r(x, y)$, όπου $C \in CN$ ένα όνομα έννοιας, $r \in RN$ ένα όνομα ρόλου και x, y μεταβλητές. Να σημειώσουμε ότι τα ατομικά ερωτήματα, μπορούν να μην εμπλέκουν μόνο μεταβλητές αλλά και σταθερές, δηλαδή κάποια από τα x, y μπορούν να είναι ονόματα ατόμων. Στην περίπτωση που το ερώτημα δεν εμπλέκει καθόλου μεταβλητές θα ονομάζεται ερώτημα αληθοτιμής ή ερώτημα Boole (boolean query).

Θα λέμε ότι η x είναι η μεταβλητή του ατομικού ερωτήματος q_1 (αντίστοιχα οι x, y είναι οι μεταβλητές του q_2), και θα αναφερόμαστε σε αυτή (αυτές) μέσω της συνάρτησης var. Δηλαδή $var(q_1) = \{x\}$ και $var(q_2) = \{x, y\}$.

Θα καλούμε συζευκτικό ερώτημα (conjunctive query) μια έκφραση της μορφής $q(x) = \{q_1, ..., q_n\}$, όπου q_i, $i \in \mathbb{N}_n$, ένα ατομικό ερώτημα. Το διάνυσμα x ονομάζεται διάνυσμα μεταβλητών (variable vector) του q, και κάθε στοιχείο του (δηλαδή κάθε μεταβλητή του q) απαιτείται να περιλαμβάνεται τουλάχιστον σε ένα ατομικό ερώτημα q_i.

Παρόλο που το πρόβλημα της απάντησης ερωτημάτων γενικά είναι ένα πολύπλοκο ζήτημα, στο οποίο εμπλέκονται έννοιες όπως η ικανοποιησιμότητα, η υπαγωγή, η συνεπαγωγή, η συνέπεια και άλλες, η κατάσταση απλοποιείται σε μεγάλο βαθμό στην περίπτωση που η γλώσσα που χρησιμοποιούμε είναι απλή και τα ερωτήματα που καλούμαστε να απαντήσουμε προκύπτουν από σύζευξη εννοιών του Abox της οντολογίας μας. Για παράδειγμα, ας θεωρήσουμε τη βάση γνώσης K με το παρακάτω Abox και κενό Tbox:

$$\alpha_1.\ Movie(Die\ Hard)$$
$$a_2.\ Movie(Die\ Hard\ 2)$$
$$a_3.\ Movie(Titanic)$$
$$a_4.\ Action(Die\ Hard)$$
$$a_5.\ Action(Die\ Hard\ 2)$$
$$a_6.\ Drama(Titanic)$$
$$a_7.\ Drama(The\ Story\ of\ Us)$$
$$a_8.\ hasProtagonist(The\ Story\ of\ Us, Bruce\ Willis)$$
$$a_9.\ hasProtagonist(Die\ Hard, Bruce\ Willis)$$
$$a_{10}.\ hasProtagonist(Die\ Hard\ 2, Bruce\ Willis)$$
$$a_{11}.\ hasProtagonist(Titanic, Kate\ Winslet)$$

Τότε, αν θέλουμε να βρούμε όλες τις ταινίες που είναι ταινίες δράσης και έχουν ως πρωταγωνιστή τον Bruce Willis, αρκεί να σχηματίσουμε το παρακάτω συζευκτικό ερώτημα:

$$q(x) = [Movie(x), Action(x), hasProtagonist(x, Bruce\ Willis)]$$

Μετά από την εκτέλεση του παραπάνω querry στη βάση γνώσης θα λάβουμε ως απαντήσεις ότι οι ταινίες δράσης που έχουν πρωταγωνιστή τον Bruce Willis είναι το Die Hard και το Die Hard 2. Στο παρακάτω παράδειγμα Abox θα προσπαθήσουμε να βρούμε το σύνολο των ταινιών που είναι ταυτόχρονα και ταινίες δράσης και ταινίες επιστημονικής φαντασίας:

$$\alpha_1.\ Action(John\ Wick)$$
$$a_2.\ Action(Star\ Wars\ Episode\ V: The\ Empire\ Strikes\ Back)$$
$$a_3.\ Action(Die\ Hard)$$
$$a_4.\ Action(Die\ Hard\ 2)$$
$$a_5.\ Drama(Titanic)$$
$$a_6.\ Drama(Forrest\ Gump)$$
$$a_7.\ SciFi(Star\ Wars\ Episode\ V: The\ Empire\ Strikes\ Back)$$
$$a_8.\ Action(The\ Expendables\ 3)$$
$$a_9.\ Animation(Shrek)$$

Σχηματίζουμε το querry: $q(x) = [Action(x), SciFi(x)]$ και μας επιστρέφεται ως μόνη απάντηση η ταινία Star Wars Episode V: The Empire Strikes Back.

3 Δημιουργία της Κινηματογραφικής Οντολογίας

Περίληψη. Σε αυτό το κεφάλαιο παρουσιάζονται οι πηγές της κινηματογραφικής οντολογίας, η διαχείριση αυτών, καθώς και ο τρόπος με τον οποίο επεξεργάστηκαν προκειμένου να αποκτήσουν δομή κατάλληλη για χρήση από τη μηχανή συστάσεων. Επίσης, κρίθηκε σκόπιμη η ανάλυση του τρόπου με τον οποίο αναπαρίσταται η πληροφορία στον παγκόσμιο ιστό και αντίστοιχα πώς αυτή η αναπαράσταση εκφράστηκε στα αρχεία που αποτέλεσαν την οντολογία.

3.1 Πηγές Δεδομένων

Προτού προβούμε στην ανάλυση της ίδιας της μηχανής συστάσεων είναι σκόπιμο να αναφερθούμε στον τρόπο με τον οποίο δομήθηκε η οντολογία από την οποία αντλούνται οι κινηματογραφικές συστάσεις.

Οι δύο βάσεις δεδομένων που χρησιμοποιήθηκαν για τη σύνταξη της κινηματογραφικής οντολογίας ήταν η βάση δεδομένων Movielens[4] και μέρος της βάσης δεδομένων του IMDB[5]. Το Movielens είναι ένα online σύστημα συστάσεων συνεργατικού φιλτραρίσματος, το οποίο δημιουργήθηκε για ακαδημαϊκούς και γενικότερους εκπαιδευτικούς λόγους. Ως εκ τούτου, το σύνολο των δεδομένων που περιέχει από προηγούμενες αλληλεπιδράσεις των χρηστών είναι διαθέσιμο για χρήση από οποιονδήποτε ιδιώτη εφόσον φυσικά γίνει η απαραίτητη αναφορά στην πηγή των πληροφοριών. Το Movielens παρέχει δύο κατηγορίες dataset, μία μικρή κατηγορία που χρησιμοποιείται για τον έλεγχο της λειτουργίας μηχανών συστάσεων και μία δεύτερη μεγάλη κατηγορία που χρησιμοποιείται για εμπορικές εφαρμογές και για ερευνητικούς σκοπούς. Πιο συγκεκριμένα, η μικρή κατηγορία περιέχει εκατό χιλιάδες βαθμολογίες ταινιών που εφαρμόστηκαν πάνω σε δέκα χιλιάδες ταινίες από εφτιακόσιους χρήστες. Η μεγάλη οντολογία περιέχει εικοσιτέσσερα εκατομμύρια βαθμολογίες ταινιών που εφαρμόστηκαν σε σαράντα χιλιάδες ταινίες από συνολικά διακόσους εξήντα χιλιάδες χρήστες. Από τα δύο σετ επιλέχτηκε το πρώτο, καθώς χαρακτηρίζεται ως πιο “σταθερό” από τους ίδιους τους δημιουργούς του Movielens και συνεπώς θεωρήθηκε

καταλληλότερο για τον έλεγχο της ορθής λειτουργίας της μηχανής συστάσεων που επρόκειτο να κατασκευάσουμε.

Εκτός από τα ratings για κάθε μία από τις δέκα χιλιάδες ταινίες, το ML Small Dataset παρείχε ακόμα links για κάθε ταινία στο tMDB (the Movie Database) και στο IMDB (Internet Movie Database), τα είδη στα οποία ανήκει κάθε ταινία και κάποιες δημοφιλείς ετικέτες τις οποίες χρησιμοποιούσαν οι χρήστες για να περιγράψουν την κάθε ταινία. Στα πλαίσια της δημιουργίας ενός συστήματος σύστασης βασισμένου σε γνώση, τα υποψήφια προς αξιοποίηση στοιχεία ήταν τα είδη στα οποία ανήκει κάθε ταινία και οι ετικέτες που έχουν δώσει οι χρήστες. Από αυτές τις δύο κατηγορίες δεδομένων, χρησιμοποιήθηκαν τα είδη των ταινιών ως έννοιες της οντολογίας, ενώ οι ετικέτες απορρίφθηκαν εξ ολοκλήρου διότι η σημασιολογία της συντριπτικής πλειονότητας αυτών αντικατόπτριζε υποκειμενικές κρίσεις για την κάθε ταινία και δεν αποτελούσαν έτσι πετυχημένες οντολογικές περιγραφές. Για παράδειγμα, έννοιες όπως “dull”, “boring”, “weird”, “Asian”, “predictable”, “emotional”, “trilogy” και άλλες μόνο να βλάψουν την ποιότητα της οντολογικής αναπαράστασης μπορούσαν. Παρόλα αυτά, ετικέτες όπως “Quentin Tarantino”, “Steve Martin”, “Indie” ή “B Movie”, δηλαδή ετικέτες που αναφέρονταν στα αντικειμενικά χαρακτηριστικά των ταινιών (συντελεστές και σκηνοθεσία), δυνητικά θα μπορούσαν να φανούν χρήσιμες, αλλά δεν υπήρχε αυτοματοποιημένη μέθοδος για τον εντοπισμό τους ανάμεσα στις μη χρήσιμες ετικέτες.

Προκειμένου να αξιοποιηθούν έννοιες παρόμοιες με αυτές που απορρίφθηκαν στο Movielens και να εμπλουτιστεί περαιτέρω η οντολογία, απαιτούνταν να αξιοποιηθούν και άλλες βάσεις δεδομένων. Από την πληθώρα κινηματογραφικών βάσεων που είναι διαθέσιμες στο διαδίκτυο, οι τρεις που ξεχώρισαν ήταν οι βάσεις του IMDB, του tMDB και του DBTropes. Το IMDB είναι η μεγαλύτερη online βάση δεδομένων που σχετίζεται με τους τομείς του κινηματογράφου και της τηλεόρασης. Στα αρχεία της υπάρχουν αποθηκευμένες πληροφορίες για 4.389.431 ταινίες και τηλεοπτικές σειρές, μαζί με τους ηθοποιούς, τους σκηνοθέτες και τους λοιπούς συντελεστές κάθε μίας από αυτές, βαθμολογίες κριτικών και χρηστών, όπως επίσης και τα χαρακτηριστικά τους, δηλαδή είδη, διάρκεια, χώρα προέλευσης, γλώσσα και δεκάδες ακόμα πληροφορίες. Η πληρότητα και η εγκυρότητα των δεδομένων

του IMDB καθιστούν τη συγκεκριμένη βάση την πλέον χρήσιμη για τη δημιουργία οποιουδήποτε συστήματος σύστασης κινηματογραφικής φύσεως. Η βάση του tMDB δεν έχει την ίδια εγκυρότητα και βάθος ώστε να μπορέσει να συγκριθεί με του IMDB λόγω του γεγονότος ότι βασίζεται σε συνεισφορές της κοινότητας για να εμπλουτιστεί, ωστόσο προσφέρει ένα εύχρηστο API το οποίο μπορεί να χρησιμοποιηθεί για τη λήψη στοιχείων μετά την παρουσίαση των αποτελεσμάτων. Τέλος, το DBTropes, η βάση δεδομένων του TVTropes, περιέχει μία πληθώρα από ετικέτες για τις περισσότερες ταινίες που έχουν κυκλοφορήσει από το έτος σύστασης του TVTropes το 2008.

3.2 Εξαγόμενες Έννοιες

Με βάση όλα τα παραπάνω επιλέχτηκαν να αντληθούν ως στοιχεία για την κινηματογραφική οντολόγια τα παρακάτω:

- Όλες οι ταινίες και τα είδη τους από το ML Small Dataset.
- Για κάθε ταινία που λάβαμε από το ML Small Dataset ελέγχαμε στη βάση του IMDB τους ηθοποιούς της. Κατόπιν, κατατάσσαμε τους ηθοποιούς ανάλογα με τον αριθμό των ταινιών που έχουν παίξει. Για κάθε ταινία κρατούσαμε τους τέσσερις ηθοποιούς που είχαν παίξει στις περισσότερες ταινίες και θεωρούσαμε ότι αυτοί οι ηθοποιοί χαρακτηρίζουν τη συγκεκριμένη ταινία.

Η διαδικασία αυτή πραγματοποιήθηκε με αυτοματοποιημένη ανάλυση των .csv αρχείων που παρέχονταν από το Movielens και το IMDB. Για κάθε ένα από τα είδη δημιουργήθηκε μία έννοια, όπως επίσης και για κάθε ηθοποιό. Μία ταινία x άνηκε στην έννοια κάποιου είδους, έστω E, αν και μόνο αν η συγκεκριμένη ταινία είχε αυτό το είδος (συμβολίζουμε με E(x)), ενώ κάθε ταινία άνηκε στην έννοια κάποιου ηθοποιού, έστω H, αν και μόνο αν ο ηθοποιός αυτός άνηκε στους τέσσερις πιο δημοφιλής ηθοποιούς της ταινίας (συμβολίζουμε με H(x)). Έτσι, για τα είδη προέκυψαν οι εξής έννοιες:

Adventure	Animation
Children	Comedy
Fantasy	Romance

Drama	Action
Crime	Thriller
Horror	Mystery
SciFi	Documentary
IMAX	War
Musical	Western
FilmNoir	Nogenreslisted

Με τον ίδιο τρόπο καταγράφηκαν και οι ηθοποιοί για κάθε ταινία. Η πλήρης λίστα με τις έννοιες που προκύπτουν από τους ηθοποιούς φαίνεται στο παράρτημα Α.

Έτσι, για παράδειγμα για την ταινία Die Hard σημειώθηκαν τα εξής: Action(Die Hard), Crime(Die Hard), Thriller(Die Hard), BruceWillis(Die Hard), BonnieBedelia(Die Hard), PaulGleason(Die Hard), AlanRickman(Die Hard).

Θα πρέπει να σημειωθεί ότι στην πραγματικότητα ο τρόπος με τον οποίο καταγράφονταν τα αντικείμενα/άτομα και δηλώνονταν οι έννοιες στις οποίες ανήκουν δεν πραγματοποιήθηκε με τον τυπικό τρόπο που έχει παρουσιαστεί μέχρι στιγμής, δηλαδή με το όνομα της έννοιας ακολουθούμενο από παρένθεση μέσα στην οποία βρίσκεται το όνομα του αντικειμένου. Αντίθετα χρησιμοποιήθηκε διαφορετική σύνταξη, η οποία σχετίζεται άμεσα με τον τρόπο με τον οποίο αναπαρίσταται η γνώση στον παγκόσμιο ιστό. Επομένως, σε αυτό το σημείο θεωρείται σκόπιμο να παρουσιαστεί ο τρόπος με τον οποίο η γνώση κωδικοποιείται ώστε να μπορεί να χρησιμοποιηθεί απρόσκοπτα από ανθρώπους και εφαρμογές.

3.3 Αναπαράσταση Γνώσης στον Παγκόσμιο Ιστό[6]

Στον Παγκόσμιο Ιστό, κατά την αναπαράσταση της πληροφορίας, γίνεται αναφορά σε ένα σύνολο αντικειμένων, προσβάσιμων μέσω του διαδικτύου, όπως ιστοτόπους, ψηφιακές εικόνες, ηλεκτρονικές διευθύνσεις, διαδικτυακές υπηρεσίες κλπ, αλλά και σε άλλα αντικείμενα του κόσμου (απτά ή αφηρημένα) που δεν είναι προσβάσιμα μέσω του διαδικτύου, όπως ανθρώπους, πόλεις, τα βιβλία μιας βιβλιοθήκης, τις ηλεκτρικές συσκευές ενός σπιτιού, μια προβολή κάποιας κινηματογραφικής ταινίας, κάποια αριθμητική τιμή κλπ. Για το σκοπό αυτό, έχει προτυποποιηθεί (μέσω της W3C) και χρησιμοποιείται στην πράξη

ένας ενιαίος τρόπος αναφοράς στα αντικείμενα του κόσμου που αποτελούν τα στοιχεία αναφοράς.

Τυπικά, τα στοιχεία αναφοράς στον Παγκόσμιο Ιστό ονομάζονται πόροι (resources). Οι πόροι ταυτοποιούνται μέσω ενός προσδιοριστικού (identifier), το οποίο για να είναι μοναδικό σε όλο τον Παγκόσμιο Ιστό, αλλά και να ερμηνεύεται εύκολα, είναι απαραίτητο να είναι ομοιόμορφο (uniform). Τα προσδιοριστικά που χρησιμοποιούνται για τον σκοπό αυτό ονομάζονται URI (Uniform Resource Identifier) και η τυπική σύνταξή τους καθορίζεται στην κοινότητα του Παγκοσμίου Ιστού από ένα πρωτόκολλο που αποτελεί μέρος του Internet Official Protocol Standards, που εκδίδεται από την W3C.

Όπως κάθε προσδιοριστικό, έτσι και τα URI έχουν ως σκοπό, μέσω του ονόματός τους, να ταυτοποιήσουν το αντικείμενο του κόσμου το οποίο προσδιορίζουν. Δηλαδή, να διακρίνουν τον ένα πόρο από τον άλλο, ανεξάρτητα από το είδος ή την προέλευσή του (αν, για παράδειγμα, είναι εικόνα, ηλεκτρονική διεύθυνση ή βιβλίο σε μια βιβλιοθήκη). Συνεπώς, αν σκεφτούμε πόσα διαφορετικά είδη αντικειμένων, προελεύσεις, κατηγορίες κοκ. υπάρχουν στον κόσμο, το όνομα ενός URI πρέπει να ενσωματώνει αρκετά στοιχεία διάκρισης, αυτά που είναι απαραίτητα για τον καθορισμό της ταυτότητας των στοιχείων αναφοράς.

Επιπλέον, είναι σημαντικό να μην υπάρχει διαφορετικός τρόπος και διαφορετικοί συντακτικοί κανόνες αναπαράστασης για κάθε τύπο προσδιοριστικού, αλλά να διατηρείται η ομοιομορφία. Με τον τρόπο αυτό υποστηρίζεται η ομοιόμορφη σημασιολογική ερμηνεία των προσδιοριστικών, καθώς και η απρόσκοπτη επέκταση των τύπων προσδιοριστικών (που πηγάζει από την ανάγκη για αναφορά σε νέα αντικείμενα του κόσμου), χωρίς να τροποποιούνται οι υπάρχοντες κανόνες. Τέλος, με τον τρόπο αυτό δίνεται η δυνατότητα για επαναχρησιμοποίηση των προσδιοριστικών και από άλλες εφαρμογές και υπηρεσίες, εκτός από αυτές που τα έχουν εισαγάγει.

Τυπικά, ένα URI είναι μια ακολουθία χαρακτήρων που χωρίζεται σε μέρη που καθορίζουν συγκεκριμένα χαρακτηριστικά του προσδιοριζόμενου πόρου. Τέτοια χαρακτηριστικά είναι το σχήμα (scheme), που καθορίζει τον τύπο του πόρου και, ανάλογα με τον τύπο, τους ιδιαίτερους συντακτικούς κανόνες σύνταξης του URI, η αρχή κατονομασίας (naming authority), που προσδιορίζει

(ανάλογα με το σχήμα) τον υπεύθυνο απόδοσης του ονόματος με βάση την αναφορά που ακολουθεί, η διαδρομή (path), που προσδιορίζει τη συγκεκριμένη αναφορά στο πλαίσιο της συγκεκριμένης αρχής, το ερώτημα (query), που προσδιορίζει τον πόρο αναφοράς μεταξύ όλων των πόρων που βρίσκονται στην εμβέλεια της αρχής και στη συγκεκριμένη διαδρομή (ανάλογα με το σχήμα), και το απόσπασμα (fragment), που προσδιορίζει μια έμμεση αναφορά σε ένα πόρο που αποτελεί τμήμα αυτού που καθορίστηκε μέχρι το ερώτημα (για παράδειγμα ένα συγκεκριμένο σημείο μέσα σε ένα αρχείο).

Έτσι για παράδειγμα το URI "**http://ntua.gr/Stefanos/DieHard**" προσδιορίζει ένα συγκεκριμένο πόρο, σε αυτή την περίπτωση την ταινία Die Hard στον Παγκόσμιο Ιστό.

Κυρίως λόγω των μακροσκελών ονομάτων που χρησιμοποιούνται για την απόδοση όλων των απαραίτητων για την ταυτοποίηση πληροφοριών στα URI, είναι ιδιαίτερα χρήσιμη μια ειδική περίπτωση ονοματοχώρων που έχει ως στόχο τον σαφή καθορισμό προθεμάτων, ώστε να καθίσταται ευκολότερη η ορθή συγγραφή και ανάγνωση URI (από τον άνθρωπο), όπου και αν αυτά χρησιμοποιούνται. Η διαδικασία δήλωσης συντμήσεων προθεματικής χρήσης αυτού του τύπου γίνεται μέσω των *ονοματοχώρων XML* (XML namespaces). Για παράδειγμα, ορίζοντας σε κάποιον ονοματοχώρο XML τον πόρο "**http://ntua.gr/Stefanos/**" ως "kbrs", θα μπορούσαμε να αναφερθούμε στον πόρο που περιγράφει την ταινία Die Hard ως "kbrs:DieHard".

Από τους ονοματοχώρους που χρησιμοποιούνται για τη σύνταξη οντολογιών στον παγκόσμιο ιστό, οι πιο χρήσιμοι είναι οι εξής:

- rdf: το οποίο καλύπτει το URI **http://www.w3.org/1999/02/22-rdf-syntax-ns#**
- rdfs: το οποίο καλύπτει το URI **http://www.w3.org/2000/01/rdf-schema#**
- owl: το οποίο καλύπτει το URI **http://www.w3.org/2002/07/owl#**
- dc: το οποίο καλύπτει το URI **http://dublincore.org/documents/dcmi-namespace/**
- foaf: το οποίο καλύπτει το URI **http://xmlns.com/foaf/0.1/.**

Σε κάθε μία από τις παραπάνω διευθύνσεις μπορεί κανείς να βρει μία οντολογία που περιγράφει ένα συγκεκριμένο λεξιλόγιο. Οι τρεις πρώτοι ονοματοχώροι χρησιμοποιούνται για να περιγράψουν κατά βάση χαρακτηριστικά και σχέσεις μεταξύ αντικειμένων, ο ονοματοχώρος DC (Dublin Core Initiative) χρησιμοποιείται για την περιγραφή ψηφιακών βιβλιοθηκών, ενώ ο foaf (friend of a friend) στοχεύει στη δημιουργία ενός κοινωνικού δικτύου σελίδων που περιγράφουν ανθρώπους, τους δεσμούς μεταξύ τους και τα πράγματα που δημιουργούν και κάνουν.

Από τους παραπάνω ονοματοχώρους, για την κατασκευή της κινηματογραφικής οντολογίας αξιοποιήσαμε τους εξής πόρους:

1. owl:Class για να δηλώσουμε τις έννοιες της οντολογίας
2. rdf:about για να δηλώσουμε τον πόρο που σχετίζεται με την κάθε έννοια
3. rdfs:label για να δηλώσουμε το όνομα της κάθε έννοιας
4. rdf:type για να δηλώσουμε ότι ένα αντικείμενο με συγκεκριμένο URI ανήκει σε μία έννοια με άλλο, επίσης συγκεκριμένο, URI

3.4 Σύνταξη της Κινημαγραφικής Οντολογίας

Με βάση λοιπόν τους παραπάνω πόρους και αντληθέντα από τις βάσεις δεδομένων στοιχεία, συντάχθηκαν τα αρχεία της οντολογίας, τα οποία χωρίζονταν σε δύο κατηγορίες:

- Αρχεία OWL (Web Ontology Language[7]) τα οποία περιείχαν τους ορισμούς των κλάσεων και
- Αρχεία NT (ontology[2]) τα οποία δήλωναν τα αντικείμενα που ανήκαν σε κάθε κλάση.

Τα αρχεία OWL είχαν την παρακάτω μορφή:

```
<owl:Class rdf:about="http://ntua.gr/Stefanos/Action">
        <rdfs:label>Action</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://ntua.gr/Stefanos/Adventure">
        <rdfs:label>Adventure</rdfs:label>
```

```
</owl:Class>
<owl:Class rdf:about="http://ntua.gr/Stefanos/BruceWillis">
        <rdfs:label>BruceWillis</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://ntua.gr/Stefanos/JasonStatham">
        <rdfs:label>JasonStatham</rdfs:label>
</owl:Class>


. . . . . .
```

Στο παραπάνω πλαίσιο απεικονίζονται ανά τρεις σειρές οι δηλώσεις των διαφόρων εννοιών της οντολογίας γραμμένες σύμφωνα με έναν ονοματοχώρο XML (xml namespace). Κάθε τριάδα γραμμών ορίζεται από την ετικέτα (tag) <owl:Class> η οποία περιέχει ως χαρακτηριστικό (attribute) τον πόρο rdf:about, ο οποίος χρησιμοποιείται για να δηλώσουμε το URI με το οποίο συνδέεται η συγκεκριμένη έννοια. Στο περιεχόμενο της ετικέτας βρίσκεται η ετικέτα-πόρος rdfs:label, που χρησιμοποιείται για την προσθήκη ενός ονόματος στη συγκεκριμένη έννοια. Το όνομα αυτό δεν έχει από μόνο του σημασιολογική αξία και υπάρχει μόνο για να προσδίδει μία φιλικότερη προς τον άνθρωπο περιγραφή της έννοιας. Σε αυτά τα πλαίσια, θα μπορούσε να παραλειφθεί και η δήλωσεις να έχουν την ακόλουθη μορφή:

```
<owl:Class rdf:about="http://ntua.gr/Stefanos/Action"/>
<owl:Class rdf:about="http://ntua.gr/Stefanos/Adventure"/>
<owl:Class rdf:about="http://ntua.gr/Stefanos/BruceWillis"/>
<owl:Class rdf:about="http://ntua.gr/Stefanos/JasonStatham"/>
. . . . . .
```

Παρόλα αυτά, γενικά αποτελεί καλή πρακτική να δίνεται μία περιγραφή σε κάθε έννοια που δηλώνεται, καθώς το URI μπορεί να είναι πολύ πιο πολύπλοκο από τις περιπτώσεις με τις οποίες ασχοληθήκαμε στην παρούσα εργασία.

Η μορφή των αρχείων NT ήταν λίγο διαφορετική από αυτή των αρχείων OWL με την έννοια ότι, ενώ διατηρούσαν τη βασική δομή των τριάδων, ο τρόπος διάρθρωσης αυτών ήταν εννοιολογικά διαφορετικός. Πιο συγκεκριμένα, το αρχείο NT αποτελούνταν από διαδοχικές τριάδες καθεμία από τις οποίες είχε ως πρώτο όρισμα τα URI των αντικειμένων (των ταινιών στην προκειμένη περίπτωση), ως δεύτερο όρισμα τον πόρο rdf:type και ως τρίτο όρισμα μίας

από τις έννοιες που έχουν οριστεί στο αρχείο OWL. Σχηματικά, οι τριάδες είχαν τη μορφή:

$$URI\ Αντικειμένου \quad rdf:type \quad URI\ Έννοιας.$$

Έχοντας υπόψη μας αυτό το γενικό πρότυπο, μπορούμε να αντιληφθούμε καλύτερα τον τρόπο με τον οποίο δομείται το αρχείο NT που αποτέλεσε τη βασική συνιστώσα της κινηματογραφικής οντολογίας. Παρακάτω φαίνεται απόσπασμα ενός από τα αρχεία NT που χρησιμοποιήθηκαν κατά τη σύσταση της οντολογίας.

```
<http://ntua.gr/Stefanos/ToyStory>      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Adventure>     .
<http://ntua.gr/Stefanos/ToyStory>      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Animation>     .
<http://ntua.gr/Stefanos/ToyStory>      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Children>      .
<http://ntua.gr/Stefanos/ToyStory>      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/ToyStory>      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Fantasy>       .
<http://ntua.gr/Stefanos/Jumanji>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Adventure>     .
<http://ntua.gr/Stefanos/Jumanji>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Children>      .
<http://ntua.gr/Stefanos/Jumanji>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>      <http://ntua.gr/Stefanos/Fantasy>       .
<http://ntua.gr/Stefanos/GrumpierOldMen>        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/GrumpierOldMen>        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Romance>       .
<http://ntua.gr/Stefanos/WaitingtoExhale>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/WaitingtoExhale>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Drama> .
<http://ntua.gr/Stefanos/WaitingtoExhale>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Romance>       .
<http://ntua.gr/Stefanos/FatheroftheBridePartII>        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/Heat>  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Action>        .
<http://ntua.gr/Stefanos/Heat>  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Crime> .
<http://ntua.gr/Stefanos/Heat>  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Thriller>      .
<http://ntua.gr/Stefanos/Sabrina>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/Sabrina>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Romance>       .
<http://ntua.gr/Stefanos/TomandHuck>    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Adventure>     .
<http://ntua.gr/Stefanos/TomandHuck>    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Children>      .
<http://ntua.gr/Stefanos/SuddenDeath>   <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Action>        .
<http://ntua.gr/Stefanos/GoldenEye>     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Action>        .
<http://ntua.gr/Stefanos/GoldenEye>     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Adventure>     .
<http://ntua.gr/Stefanos/GoldenEye>     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Thriller>      .
<http://ntua.gr/Stefanos/AmericanPresident>     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/AmericanPresident>     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Drama> .
<http://ntua.gr/Stefanos/AmericanPresident>     <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Romance>       .
<http://ntua.gr/Stefanos/DraculaDeadandLovingIt>        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Comedy>        .
<http://ntua.gr/Stefanos/DraculaDeadandLovingIt>        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Horror>        .
<http://ntua.gr/Stefanos/Balto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Adventure>     .
<http://ntua.gr/Stefanos/Balto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Animation>     .
<http://ntua.gr/Stefanos/Balto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Children>      .
<http://ntua.gr/Stefanos/Nixon> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Drama> .
<http://ntua.gr/Stefanos/CutthroatIsland>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Action>        .
<http://ntua.gr/Stefanos/CutthroatIsland>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Adventure>     .
<http://ntua.gr/Stefanos/CutthroatIsland>       <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://ntua.gr/Stefanos/Romance>       .
```

Όσον αφορά την υλοποίηση της αυτοματοποιημένης δημιουργίας της οντολογίας, επιλέχτηκε η γλώσσα C#, ενώ τα προγράμματα αναπτύχθηκαν στο Visual Studio 2015. Η επιλογή αυτή έγινε λόγω των ενσωματωμένων εργαλείων που παρέχει το Visual Studio, όπως το LINQ, το ADO.NET και το Entity Framework[8], τα οποία δίνουν τη δυνατότητα ταχύτερης και αποτελεσματικότερης διαχείρισης δεδομένων που πηγάζουν από βάσεις δεδομένων (όπως σε αυτή την περίπτωση τα .csv αρχεία του Movielens και του IMDB). Πιο συγκεκριμένα, τα εργαλεία αυτά μετατρέπουν δεδομένα μεταξύ ασύμβατων τύπων σε αντικειμενοστραφείς περιγραφές, εύκολα διαχειρίσιμες από γλώσσες όπως η C# και η Java. Η διαδικασία αυτή καλείται Αντικειμενοστραφής Σχεσιακή Χαρτογράφηση (Object Relational Mapping, ORM) και επιτάχυνε ιδιαίτερα τη διαδικασία ανάπτυξης της οντολογίας.

4 Λειτουργία της Μηχανής Συστάσεων

Περίληψη. Το παρόν κεφάλαιο περιγράφει σε βάθος τις αρχές λειτουργίας της μηχανής συστάσεων όταν ο χρήστης έχει δηλώσει ένα αντικείμενο του χώρου ενδιαφέροντος που τον αφορά. Αναλύονται ζητήματα από την επεξεργασία της οντολογίας έως την παρουσίαση των αποτελεσμάτων.

4.1 Δημιουργία του Οντολογικού Γράφου[9]

Όπως αναφέρθηκε στα προηγούμενα κεφάλαια, στόχος των συστήματων σύστασης αποτελεί η παροχή προτάσεων στο χρήστη οι οποίες χαρακτηρίζονται από σχετικότητα, καινοτομία, εναλλακτικότητα και ποικιλομορφία. Το σύστημα σύστασης βασισμένο σε γνώση που σχεδιάστηκε επιδίωξε να πετύχει και τους τέσσερις αυτούς στόχους μέσω της κατάλληλης σχεδίασης της λειτουργίας της μηχανής συστάσεων. Ο βασικός πυλώνας στον οποίο στηρίζεται αυτή η σχεδίαση είναι η αξιοποίηση της κινηματογραφικής οντολογίας, η δημιουργία της οποίας περιγράφηκε στο προηγούμενο κεφάλαιο. Πιο συγκεκριμένα, οι συστάσεις που παρουσιάζονται στο χρήστη αποτελούν αντικείμενα της οντολογίας, ενώ η λογική των συστάσεων συνδέεται άμεσα με τις έννοιες στις οποίες ανήκει κάθε αντικείμενο. Σε γενικές γραμμές, για κάθε αντικείμενο το οποίο δηλώνει ο χρήστης ότι τον ενδιαφέρει, γίνεται ανάλυση των εννοιών στις οποίες ανήκει και παρουσιάζονται αντικείμενα των οποίων το σύνολο εννοιών είναι το ίδιο με αυτό του αντικειμένου που ενδιαφέρει το χρήστη ή λίγο διαφορετικό (για παράδειγμα διαφέρει ως προς μία ή δύο έννοιες).

Προκειμένου να μπορεί να γίνει άμεση εκτίμηση των εννοιών που ανήκει ένα αντικείμενο, όπως και εύρεση των αντικειμένων που μοιάζουν με αυτό, απαιτούνταν η μετατροπή της οντολογίας σε μία δομή εύκολα προσβάσιμη και υλοποιήσιμη σε μία αντικειμενοστραφή γλώσσα, όπως η Java και η C#. Για αυτό το σκοπό επιλέχτηκε η αξιοποίηση της δομής ενός ακυκλικού κατευθυνόμενου γράφου (directed acyclic graph, DAG).

Πιο συγκεκριμένα, βασική δομική μονάδα του DAG είναι ο κόμβος, ο οποίος περιέχει μία ετικέτα και ένα σύνολο αντικειμένων. Για κάθε ετικέτα, η οποία αποτελεί υποσύνολο του συνόλου των εννοιών που έχουν παραχθεί, το

αντίστοιχο σύνολο αντικειμένων αποτελείται από όλα τα αντικείμενα που ανήκουν στις έννοιες της ετικέτας. Το γεγονός αυτό επιτυγχάνεται σε δύο βήματα. Κατά το πρώτο βήμα, δημιουργείται ένας αρχικός DAG όπου οι ετικέτες κάθε κόμβου του είναι μονοσύνολα. Αντίστοιχα συμπληρώνονται και τα σύνολα αντικειμένων του. Αν κάποιος κόμβος περιέχει σύνολο αντικειμένων που είναι υποσύνολο κάποιου άλλου κόμβου, τότε αυτός ο κόμβος τοποθετείται ως κόμβος-παιδί του άλλου. Πάντα σε αυτό το γράφο υπάρχει ένας κόμβος με ετικέτα Top ο οποίος περιέχει όλα τα αντικείμενα. Μετά τον ολοκλήρωση αυτής της διαδικασίας εκκινεί δεύτερη διαδικασία όπου ο γράφος μετασχηματίζεται και παίρνει την τελική του μορφή. Οι κόμβοι παιδιά του κόμβου root (όλοι οι κόμβοι δηλαδή) τοποθετούνται σε μία λίστα. Επαναληπτικά, πραγματοποιούμε όλες τις δυνατές τομές κόμβων τα αποτελέσματα των οποίων τοποθετούμε σε κατάλληλη ιεραρχία. Αν κάποιος συνδυασμός έχει ήδη γίνει, τότε αυτός δεν επαναλαμβάνεται.

Από τη διαδικασία δημιουργίας του γράφου γίνονται φανερά τα εξής χαρακτηριστικά:

- Μέσα στο γράφο υπάρχει ένας κόμβος του οποίου η ετικέτα είναι κενή. Σε αυτό τον κόμβο περιέχονται όλα τα αντικείμενα τα οποία ανήκουν στην έννοια owl:Thing, δηλαδή όλα τα αντικείμενα της κινηματογραφικής οντολογίας. Ο κόμβος αυτός ονομάζεται Top και περιέχει όλες τις ταινίες που εκμαιεύθηκαν από το Movielens. Πολύ σημαντικό είναι το γεγονός ότι ο κόμβος αυτός δεν έχει πατρικούς κόμβους, αλλά μόνο κόμβους παιδιά.
- Μέσα στο γράφο υπάρχει επίσης ένας κόμβος του οποίου το σύνολο των αντικειμένων είναι κενό. Η ετικέτα αυτού του κόμβου στη γενική περίπτωση δεν είναι κενή, αλλά το περιεχόμενό της δεν επηρεάζει τον τρόπο λειτουργίας του συστήματος. Ο κόμβος αυτός αποκαλείται Bottom, δεν έχει κόμβους παιδιά, ενώ κόμβοι γονείς του είναι μόνο οι κόμβοι που αν δεν υπήρχε αυτός ο κόμβος δε θα είχαν καθόλου παιδιά.
- Όσο μεγαλύτερη είναι η απόσταση ενός κόμβου από τον Top, τόσο μεγαλύτερη είναι η ετικέτα του.

- Αν ένας κόμβος Α έχει ετικέτα L, τότε κάθε κόμβος Β, του οποίου η ετικέτα είναι υπερσύνολο του L, έχει σύνολο αντικειμένων με μέγεθος το πολύ ίσο με το μέγεθος του συνόλου αντικειμένων του κόμβου Α.
- Όλα τα αντικείμενα (ταινίες) που βρίσκονται στο σύνολο αντικειμένων του ίδιου κόμβου ανήκουν στις έννοιες που υπάρχουν στην ετικέτα του συγκεκριμένου κόμβου. Έτσι για παράδειγμα, όλες οι ταινίες που περιέχονται στον κόμβο με ετικέτα [Action, Crime, Thriller, Bruce Willis], δηλαδή στην προκειμένη περίπτωση οι ταινίες Die Hard, Die Hard 2, Die Hard: With A Vengeance, Live Free or Die Hard και Good Day to Die Hard, ανήκουν στις έννοιες Action, Crime Thriller και Bruce Willis.
- Κάθε ταινία μπορεί να βρίσκεται στο σύνολο αντικειμένων παραπάνω από ενός κόμβου. Για παράδειγμα, η ταινία Red βρίσκεται και στον κόμβο με ετικέτα [Action, Comedy, Thriller, Bruce Willis] και στον κόμβο [Action, Comedy, Crime, Morgan Freeman].

Ως είσοδο η μηχανή συστάσεων λαμβάνει το γράφο που δημιουργείται από την παραπάνω διαδικασία και ένα σύνολο αντικειμένων τα οποία εισάγει ο χρήστης και αποκαλούνται σύνολο προτιμήσεων. Η βασική ιδέα πίσω από το σύστημα συστάσεων βασισμένο σε γνώση που σχεδιάστηκε είναι η εύρεση των κόμβων με τη μεγαλύτερη ετικέτα που περιέχουν τις ταινίες του συνόλου προτιμήσεων του χρήστη. Όλα τα αντικείμενα αυτών των κόμβων – εκτός από τα αντικείμενα που ανήκουν στο σύνολο προτιμήσεων – είναι υποψήφιες συστάσεις για το χρήστη. Αν ο αριθμός των υποψήφιων αυτών συστάσεων δε φτάσει ένα συγκεκριμένο αριθμό (ορίστηκε ως το τριάντα για την υλοποίηση του συστήματος), τότε στις υποψήφιες συστάσεις προστίθενται και τα σύνολα αντικειμένων των γονέων των επιλεγμένων κόμβων.

Η διαδικασία αυτή περιγράφεται αναλυτικότερα για σύνολα προτιμήσεων με ένα ή περισσότερα αντικείμενα παρακάτω.

4.2 Παραγωγή Συστάσεων από Σύνολα Προτιμήσεων Ενός Αντικειμένου

Ο αλγόριθμος για την παραγωγή συστάσεων από σύνολα προτιμήσεων ενός αντικειμένου είναι ο εξής:

1. Διέτρεξε το γράφο και βρες τον κόμβο που περιέχει το αντικείμενο του συνόλου προτιμήσεων με τη μεγαλύτερη ετικέτα.
2. Αποθήκευσε το μέγεθος της ετικέτας, έστω max.
3. Ξαναδιέτρεξε το γράφο και βρες όλους τους κόμβους που περιέχουν το αντικείμενο του συνόλου προτιμήσεων και έχουν μήκος ετικέτας ίσο με max και αποθήκευσέ τους στη λίστα initials.
4. Αρχικοποίησε ένα μετρητή συστάσεων στο 0 και μία λίστα συστάσεων στην κενή λίστα.
5. Όσο ο μετρητής συστάσεων είναι μικρότερος του 30 και υπάρχουν ακόμα αντικείμενα στους κόμβους της λίστα initials, επανάλαβε:
    1. Για κάθε κόμβο στη λίστα initials, πάρε μία σύσταση από αυτόν τον κόμβο και μετέφερέ την στη λίστα των συστάσεων.
    2. Για κάθε σύσταση που τοποθετείται στη λίστα των συστάσεων αύξησε το μετρητή συστάσεων κατά ένα.
    3. Αν μετά από οποιαδήποτε αύξηση του μετρητή συστάσεων, αυτός ξεπεράσει το 30, κόψε την επανάληψη.
6. Αν μετά την ολοκλήρωση της επανάληψης ο μετρητής συστάσεων είναι μικρότερος του 30, βρες όλους τους κόμβους γονείς των κόμβων της λίστας initials και αποθήκευσέ τους στη λίστα parents.
7. Όσο ο μετρητής συστάσεων είναι μικρότερος του 30 και υπάρχουν ακόμα αντικείμενα στους κόμβους της λίστας parents, επανάλαβε:
    1. Για κάθε κόμβο στη λίστα parents, πάρε μία σύσταση από αυτόν τον κόμβο και μετέφερέ την στη λίστα των συστάσεων.
    2. Για κάθε σύσταση που τοποθετείται στη λίστα των συστάσεων αύξησε το μετρητή συστάσεων κατά ένα.
    3. Αν μετά από οποιαδήποτε αύξηση του μετρητή συστάσεων, αυτός ξεπεράσει το 30, κόψε την επανάληψη.
8. Εμφάνισε τις συστάσεις.

Στο τέλος της εκτέλεσης του παραπάνω αλγορίθμου θα έχουν, κατά κανόνα, εμφανιστεί στο χρήστη 30 ταινίες ως συστάσεις. Για την ολοκλήρωση του αλγορίθμου θα χρειαστούν βήματα ίσα με:

$$Μέγιστος\ Αριθμός\ Βημάτων = 2*N + M + 30$$

όπου N ο αριθμός των κόμβων του γράφου και M ο αριθμός των γονέων των κόμβων της λίστας initials – κατά κανόνα μεγαλύτερος από 30. Η συγκεκριμένη υλοποίηση ενδεχομένως να φαίνεται ότι απαιτεί μεγάλο

υπολογιστικό κόστος, αλλά αν αναλογιστεί κανείς ότι ο αριθμός των κόμβων που προκύπτει από μία οντολογία με πάνω από είκοσι χιλιάδες έννοιες οδηγεί σε γράφο με μόλις εξήντα χιλιάδες κόμβους, ο χρόνος ο οποίος απαιτείται για τη διεκπεραίωση της παραπάνω διαδικασίας είναι της τάξεως των μερικών μικρών του δευτερολέπτου. Χαρακτηριστικά, για την παραγωγή συστάσεων για την ταινία Star Wars Episode V: The Empire Strikes Back, ο χρόνος που χρειάστηκε ο επεξεργαστής Intel i7 7500U ανερχόταν στα 34 μs.

Σε αυτό το σημείο αξίζει να εξεταστούν δύο παραδείγματα για να κατανοηθεί καλύτερα ο τρόπος με τον οποίο λειτουργεί ο αλγόρθμος.

Έστω λοιπόν ότι ο χρήστης ήθελε να αναζητήσει ταινίες με βάση το σύνολο προτιμήσεων που περιέχει μόνο την ταινία “Star Wars Episode V The Empire Strikes Back”.

Κατά την πρώτη αναζήτηση συστάσεων για τη συγκεκριμένη ταινία ο αλγόριθμος εντόπισε τους κόμβους N1 και N2 με ετικέτες:

- [Adventure, Hamill Mark, Action, Sci-Fi] και
- [Adventure, Action, Sci-Fi, Harisson Ford]

Τα σύνολα αντικειμένων που ανήκουν σε αυτούς τους κόμβους περιέχουν 4 και 3 συστάσεις αντίστοιχα. Οι 7 συστάσεις που παρέχονται από αυτούς τους κόμβους δεν είναι αρκετές για να συμπληρώσουν τις 30 ταινίες που έχουν τεθεί ως ο ελάχιστος αριθμός συστάσεων που παρέχονται από το σύστημα. Επομένως, το πεδίο των συστάσεων επεκτείνεται για να συμπεριληφθούν και τα σύνολα αντικειμένων των γονέων των κόμβων N1 και N2. Οι νέοι κόμβοι που προστίθενται έχουν ως ετικέτες:

- [Adventure, Hamill Mark]
- [Action, Hamill Mark]
- [Action, Sci-Fi, Hamill Mark]
- [Adventure, Sci-Fi, Hamill Mark]
- [Action, Adventure, Sci-Fi]
- [Action, Sci-Fi, Ford Harrison]

- [Action, Adventure, Hamill Mark]
- [Action, Adventure]
- [Adventure, Action, Ford Harrison]
- [Action, Ford Harrison]

Από κάθε έναν από αυτούς τους κόμβους επιλέγεται κυκλικά μία ταινία για να προστεθεί στη λίστα των συστάσεων. Όταν συμπληρωθούν 30 ταινίες, η διαδικασία διακόπτεται και οι 30 συστάσεις παρουσιάζονται στο χρήστη.

Η διαδικασία παραγωγής των συστάσεων είναι πιο απλή στην περίπτωση που ήδη από την εύρεση των αρχικών κόμβων ο αριθμός των συστάσεων καλύπτει είναι ίσος ή μεγαλύτερος του 30. Αν για παράδειγμα, εισαχθεί στο σύστημα η ταινία “Matrix”, τότε ο κόμβος που ανιχνεύεται σαν κόμβος για την εκμαίευση των συστάσεων ο κόμβος με ετικέτα [Thriller, Action, Sci-Fi]. Ο κόμβος αυτός περιέχει από μόνος του 145 ταινίες, ενώ δεν υπάρχει άλλος κόμβος με ετικέτα ίσου ή μεγαλύτερου μεγέθους. Επομένως, επιλέγονται 30 τυχαίες ταινίες από τις 145 του συγκεκριμένου κόμβου και παρουσιάζονται στο χρήστη.

Σε αυτό το σημείο με αφορμή τις ταινίες που προτείνονται ως συστάσεις για την ταινία “Star Wars Episode V: The Empire Strikes Back” θα εξεταστεί ο τρόπος με τον οποίο η λειτουργία του συστήματος πετυχαίνει τους τέσσερις βασικούς στόχους των συστημάτων σύστασης:

- Σχετικότητα: Προτείνοντας ως συστάσεις τις ταινίες οι οποίες έχουν όλα, σχεδόν όλα ή ένα μέρος από τα χαρακτηριστικά των ταινιών που δίνονται στο σύνολο προτιμήσεων διασφαλίζεται ότι οι ταινίες που συστήνονται στο χρήστη είναι σε μεγάλο βαθμό σχετικές με τις ταινίες που έχει δηλώσει ότι προτιμάει. Για παράδειγμα, προτείνονται οι ταινίες “Star Wars Episode VI: Return of the Jedi” και “Star Wars Episode VII: The Force Awakens” οι οποίες είναι οι πλέον σχετικές ταινίες με το “Star Wars Episode V”.
- Καινοτομία: Το Movielens Dataset περιέχει ακόμα και ταινίες οι οποίες δεν είχαν κυκλοφορήσει τη στιγμή της δημιουργίας του. Με αυτό τον τρόπο βεβαιωνόμαστε ότι οι συστάσεις που θα παρέχονται στους χρήστες θα περιέχουν και ταινίες οι οποίες, είτε είναι πολύ πρόσφατες, είτε δεν έχουν προβληθεί ακόμα στους κινηματογράφους. Για

παράδειγμα, παρέχεται ως σύσταση η ταινία “Spider-Man: Homecoming”, η οποία τη στιγμή της συγγραφής αυτών των γραμμών αναμενόταν να πραγματοποιήσει παγκόσμια πρεμιέρα σε μία εβδομάδα. Βέβαια, προκειμένου το σύστημα να μπορεί να συνεχίσει να επιτυγχάνει το στόχο της καινοτομίας, η βάση γνώσης που χρησιμοποιεί θα πρέπει να ενημερώνεται συχνά για τις ταινίες που πρόκειται να κυκλοφορήσουν στο μέλλον. Η διαδικασία αυτή δεν καλύπτεται στην παρούσα εργασία, αλλά γίνεται σύντομη αναφορά σε αυτήν στο τελευταίο κεφάλαιο.

- Εναλλακτικότητα: Όπως έχει αναφερθεί, η εναλλακτικότητα αποτελεί στόχο των συστημάτων σύστασης ο οποίος συνδέεται με την παροχή συστάσεων που ο χρήστης δεν περιμένε να λάβει και έτσι προκαλούν την έκπληξη και το ενδιαφέρον του. Σε κάθε περίπτωση, οι συγκεκριμένες συστάσεις δε θα πρέπει να δουλεύουν ενάντια στο στόχο της σχετικότητας και ως εκ τούτου, η ομοιότητα των προτεινόμενων αντικειμένων με τα αντικείμενα του συνόλου προτιμήσεων θα πρέπει να λαμβάνεται σοβαρά υπόψη. Η διαφορά έγγειται κυρίως στη χαλάρωση των απαιτήσεων που προκύπτουν από το σύνολο προτιμήσεων του χρήστη προκειμένου να παρουσιαστούν και άλλες ταινίες οι οποίες υπό άλλες συνθήκες θα ήταν δύσκολο να εντοπιστούν από το χρήστη. Σε αυτά τα πλαίσια, ανάμεσα στις ταινίες που προτείνονται για το “Star Wars Episode V: The Empire Strikes Back” βρίσκεται και η ταινία “Jay and Silent Bob Strike Back”, μία κωμωδία με πρωταγωνιστές τους Jason Mewes και Kevin Smith. Στη συγκεκριμένη ταινία κάνουν guest εμφανίσεις ο Mark Hamill, ο Harrison Ford και η Carrie Fischer – πρωταγωνιστές του “Star Wars Episode V” – ενώ η ταινία διαθέτει πολλές αναφορές στη σειρά ταινιών Star Wars.

- Ποικιλομορφία: Η ποικιλομορφία διασφαλίζεται από το γεγονός ότι το σύστημα επιδιώκει τη συλλογή συστάσεων οι οποίες προκύπτουν από ένα σύνολο κόμβων και απλά από έναν συγκεκριμένο (πχ τον κόμβο με τη μεγαλύτερη ετικέτα). Έτσι, για την ταινία που εξετάζεται, επιλέγονται ταινίες από 12 διαφορετικούς κόμβους, γεγονός το οποίο διασφαλίζει την ποικιλία των αποτελεσμάτων. Αυτό που πρέπει να τονιστεί είναι ότι για παράδειγμα ο κόμβος με ετικέτα [Action,

Adventure] δεν περιέχει τις ταινίες που είναι αποκλειστικά Action και Adventures, αλλά όλες τις ταινίες που ανήκουν *τουλάχιστον* στις έννοιες Action και Adventure, δηλαδή και τις ταινίες που είναι Action, Adventure, Comedy και τις ταινίες που είναι Action, Adventure, Sci-Fi κοκ.

5 Επέκταση του Συστήματος για Πολύπλοκα Σύνολα Προτιμήσεων

Μέχρι τώρα έχει περιγραφεί η λειτουργία του συστήματος όπως αυτή πραγματοποιείται σε περίπτωση που ο χρήστης επιθυμεί τη λήψη συστάσεων με βάση ενός αντικειμένου προτίμησής του. Σε αυτά τα πλαίσια, το σύστημα συστάσεων, όπως έχει ήδη αναφερθεί κατηγοριοποιεί το αντικείμενο στον κόμβο που βρίσκεται πιο κοντά στον καταληκτικό κόμβο του κατευθυνόμενου γράφου που δημιουργείται και προτείνει τα ανάλογα συγγενικά αντικείμενα. Το πρόβλημα της λήψης συστάσεων γίνεται πιο πολύπλοκο στην περίπτωση που ο χρήστης επιθυμεί τη λήψη συστάσεων, όπως αυτές προκύπτουν από ένα σύνολο αντικειμένων προτίμησης.

Στις περιπτώσεις που ο χρήστης παρέχει ομοιογενές σύνολο αντικειμένων, δηλαδή στην περίπτωση που τα αντικείμενα του συνόλου ανήκουν σε πολλές κοινές κλάσεις, το πρόβλημα αυτό μπορεί εύκολα να αντιμετωπιστεί με την εύρεση του πιο γενικού κόμβου που περιέχει και τα δύο αντικείμενα και κατόπιν με τη λήψη συστάσεων με βάση το συγκεκριμένο κόμβο. Αντίθετα, στις περιπτώσεις που το σύνολο αυτό δεν είναι αρκετά ομοιογενές, δηλαδή στις περιπτώσεις που τα αντικείμενα του συνόλου έχει λίγες ή και καμία κοινές κλάσεις, το σύστημα συστάσεων, όπως περιγράφηκε παραπάνω δυσκολεύεται να δώσει ακριβείς συστάσεις, υπονομεύοντας έτσι την κατά τα άλλα συνεπή λειτουργία του. Το πρόβλημα αυτό εμφανίζεται λόγω του γεγονότος ότι αντικείμενα με λίγες κοινές κλάσεις, όπως για παράδειγμα τα αντικείμενα “Titanic” και “Fast Furious” της κινηματογραφικής οντολογίας, ωθούν το σύστημα σε κόμβους που βρίσκονται πολύ κοντά στην αρχή του κατευθυνόμενου γράφου. Αυτό έχει ως αποτέλεσμα, οι κόμβοι των οποίων τα αντικείμενα προτείνονται να περιέχουν έως και χιλιάδες διαφορετικές προτάσεις.

5.1 Πρώτη προσέγγιση επίλυσης του προβλήματος

Προκειμένου να επιλυθεί το πρόβλημα, η βασική ιδέα που εφαρμόστηκε αρχικά σχετιζόταν με τη διαχείριση του ονομοιογενούς συνόλου ως ένωση δύο ή

περισσότερων ομοιογενών συνόλο. Η ανάγκη της εφαρμογής αυτής της αντιμετώπισης αποφασιζόταν στη διάρκεια του χρόνου εκτέλεσης στην περίπρωση που το σύστημα συνειδητοποιούσε ότι ο προτεινόμενος αριθμός ταινιών ξεπερνούσε κάποιο κατώφλι προτάσεων. Το κατώφλι προτάσεων αυτό είχε επιλεχτεί με βάση τον αριθμό των ταινιών του μεγαλύτερου κατά μέτρο κόμβου που βρισκόταν στο μεσαίο επίπεδο του Κατευθυνόμενου Γράφου. Πιο συγκεκριμένα, ο ψευδοκώδικας για τον αρχικό υπολογισμό του κατωφλίου δίνεται παρακάτω:

1. Διέτρεξε τον κατευθυνόμενο γράφο και βρες το μήκος του.
2. Βρες το σύνολο των κόμβων που απέχουν από την αρχή του γράφου $\mu\acute{\eta}\kappa o\varsigma/2$.
3. Βρες τον κόμβο με τα περισσότερα αντικείμενα.
4. Κάνε την τιμή του κατωφλίου ίση με το μέγεθος του ευρεθέντος κόμβου.

Η επιλογή του κατωφλίου εκ πρώτης όψεως μοιάζει να μην έχει σημαντική σημασιολογική βάση, ωστόσο κάτι τέτοιο δεν είναι αληθές. Όπως έχει αναλυθεί προηγουμένως, κάθε κόμβος με απόσταση x από την αρχή του κατευθυνόμενου γράφου περιέχει λιγότερα αντικείμενα όσο μεγαλύτερη είναι η απόσταση x. Σε αυτά τα πλαίσια, ο ευρεθείς κόμβος του παραπάνω αλγορίθμου βρίσκεται σε τέτοια απόσταση από την αρχή, έτσι ώστε, όλοι οι κόμβοι-πρόγονοί του να έχουν περισσότερα αντικείμενα και όλοι οι κόμβοι-απόγονοί του να έχουν λιγότερα αντικείμενα. Με αυτό τον τρόπο αποτελεί το μέσο του κατευθυνόμενου γράφου και συνεπώς αποτελεί τιμή, αρκετά σημαντική για να οριστεί ως κατώφλι στο συγκεκριμένο αλγόριθμο. Αν ένα σύνολο προτιμήσεων έχει σαν αποτέλεσμα τη λήψη προτάσεων οι οποίες σαν σύνολο είναι περισσότερες από το κατώφλι, τότε το σύνολο προτιμήσεων διασπάται σε τόσα υποσύνολα όσα ο αριθμός των προτιμήσεων και ο αλγόριθμος τρέχει ισάριθμες φορές.

Ο πρώτος, λοιπόν, αλγόριθμος που εφαρμόστηκε για την εύρεση συστάσεων που προκύπτουν από ανομοιογενή σύνολα προτιμήσεων παρουσιάζεται παρακάτω:

1. Βρες τον κόμβο που βρίσκεται πιο κοντά στο τελευταίο επίπεδο του

γράφου και περιέχει και όλα τα αντικείμενα του συνόλου προτιμήσεων

2. Αν αυτός ο κόμβος περιέχει περισσότερα αντικείμενα από το κατώφλι:
    1. Διάσπασε το σύνολο προτιμήσεων σε n διαφορετικά υποσύνολα, όπου n ο αριθμός των στοιχείων του συνόλου προτιμήσεων
    2. Βρες συστάσεις για κάθε ένα από τα διαφορετικά υποσύνολα
3. Αλλιώς παρουσίασε τα αντικείμενα αυτού του κόμβου

5.2 Επίλυση του προβλήματος με χρήση βαθμού ομοιότητας

Παρόλο που ο παραπάνω αλγόριθμος έδωσε αρκετά ικανοποιητικά αποτελέσματα στην περίπτωση συνόλου προτιμήσεων που περιείχαν δύο αντικείμενα, η αποτελεσματικότητά του έπεφτε καθώς το μέγεθος του συνόλου προτιμήσεων αυξανόταν. Το γεγονός αυτό οφειλόταν στο ότι μερικώς ομοιογενή σύνολα αντιμετωπίζονταν από τον κώδικα ως πλήρως ανομοιογενή επιστρέφοντας προτάσεις που δεν ανταποκρίνονταν στο γενικότερο σύνολο. Για παράδειγμα, το σύνολο προτιμήσεων [Titanic, Fast Furious, Fast Five] σύμφωνα με τον παραπάνω αλγόριθμο θα διασπαστεί στα σύνολα προτιμήσεων [Titanic], [Fast Furious] και [Fast Five], για καθένα από τα οποία το σύστημα συστάσεων θα προτείνει αντικείμενα ξεχωριστά. Ιδανικά, θα θέλαμε, το συγκεκριμένο σύστημα προτιμήσεων να διασπαστεί σε δύο υποσύνολα: [Titanic] και [Fast Furious, Fast Five], καθώς διαισθητικά γίνεται αντιληπτό ότι, ενώ η ταινία Titanic είναι τελείως διαφορετική από τις ταινίες Fast Furious και Fast Five, οι δύο τελευταίες ταινίες είναι σχεδόν ίδιες – ανήκουν στις ίδιες κατηγορίες ταινιών και έχουν σχεδόν τους ίδιους πρωταγωνιστές. Κατ' αναλογία με τη διαισθητική θεώρηση του ζητήματοε, η διάσπαση του συνόλου προτιμήσεων που πραγματοποιείται από το σύστημα, θα πρέπει να λαμβάνει υπόψη την ομοιότητα μεταξύ των αντικειμένων του συνόλου.

5.2.1 Σύνδεση του προβλήματος ανομοιογενούς συνόλου στιγμιοτύπων με το πρόβλημα κάλυψης συνόλου

Δοθέντος ενός συνόλου προτιμήσεων A το οποίο αποτελείται από έναν αριθμό αντικειμένων {i1, i2, i3, ..., iN}, το σύνολο των πιθανών υποσυνόλων στα οποία θα μπορούσε να χωριστεί (powerset) δίνεται ως εξής:

```
def powerset(set)
  return [set] if set.empty?


  p = set.pop
  subset = powerset(set)
  subset | subset.map { |x| x | [p] }
end
```

Τα σύνολα που προκύπτουν για έστω N = 5 είναι τα εξής:

```
[i1],[i2],[i3],[i4],[i5]
[i1,i2],[i1,i3],[i1,i4],[i1,i5]
[i1,i2,i3],[i1,i2,i4],[i1,i2,i5]
[i1,i3,i4],[i1,i3,i5],[i1,i4,i5]
[i2,i3],[i2,i4],[i2,i5]
[i2,i3,i4],[i2,i3,i5]
[i3,i4],[i3,i5]
[i3,i4,i5]
[i4,i5]
```

Σκοπός, επομένως, της δεύτερης υλοποίησης που πραγματοποιήθηκε είναι η επιλογή του κατάλληλου συνδυασμού υποσυνόλων, τα οποία καλύπτουν όλα τα αντικείμενα του συνόλου προτιμήσεων, ενώ ταυτόχρονα αποτελούν το καθένα ξεχωριστά ομοιογενές σύνολο. Το πρόβλημα αυτό μπορεί εύκολα να αντιμετωπιστεί ως το πρόβλημα της κάλυψης συνόλου, το οποίο αποδείχτηκε ότι πρόκειται για NP-hard πρόβλημα από τον Karp το 1972.

Το πρόβλημα κάλυψης συνόλου μπορεί να διατυπωθεί ως εξής:

Δοθέντος ενός συνόλου U, n αντικειμένων και ενός συνόλου υποσυνόλων του

U, έστω S = {S1, S2, ..., Sm}, όπου κάθε υποσύνολο Si έχει ένα κόστος Cost(Si), να βρεθεί το υποσύνολο του S που περιέχει όλα τα n αντικείμενα του U δίνοντας παράλληλα το μικρότερο δυνατό κόστος.

Το πρόβλημα αυτό και η αντίστοιχη απόδειξη της μη επιλυσιμότητάς του σε πολυωνυμικό χρόνο οδήγησαν στη δημιουργία των θεμελιωδών τεχνικών σε ολόκληρο το πεδίο των προσεγγιστικών αλγορίθμων. Ανάλογης σημασίας ήταν και η πληθώρα των προσεγγιστικών λύσεων που έλαβε το συγκεκριμένο πρόβλημα από το 1972 έως σήμερα. Για την επίλυση του προβλήματος όπως αυτό παρουσίαζεται στην περίπτωση της διάσπασης του συνόλου προτιμήσεων επιλέχτηκε ένας προσεγγιστικός άπληστος (greedy) αλγόριθμος, ο οποίος βασίζεται στην επαναληπτική επιλογή κάθε φορά του καλύτερου υποσυνόλου[12].

Έστω ένα σύνολο U, n αντικειμένων, ένα σύνολο υποσυνόλων του U, έστω S = {S1, S2, ..., Sm} και μία συνάρτηση κόστους Cost( ) που δρα πάνω στα στοιχεία του S.

1. Έστω I το σύνολο των αντικειμένων που έχουν εισαχθεί. Αρχικά I = { }

2. Όσο I διάφορο του U, επανάλαβε

   - Βρες το σύνολο Si με τη μικρότερη αποτελεσματικότητα κόστους, δηλαδή το πηλίκο του Cost(Si) και του αριθμού των νέων στοιχείων που προσθέτει στο I το Si: Cost(Si) / |Si – I|
   - Πρόσθεσε τα στοιχεία του Si στο I: I = I U Si

Έστω για παράδειγμα ότι έχουμε τα εξής στοιχεία:

U = {1, 2, 3, 4, 5}

S = {S1, S2, S3}

S1 = {4, 1, 3}, Cost(S1) = 5

S2 = {2, 5}, Cost(S2) = 10

S3 = {1, 4, 3, 2}, Cost(S3) = 3

Τότε, κατά την πρώτη επανάληψη έχουμε:

I = { }

Η αποτελεσματικότητα κόστους του S1 είναι ίση με 5/3.

Η αποτελεσματικότητα κόστους του S2 είση ίση με 10/2.

Η αποτελεσματικότητα κόστους του S3 είναι ίση με 3/4.

Άρα, επιλέγεται το S3 και το I γίνεται {1, 4, 3, 2}.

Κατά τη δεύτερη επανάληψη έχουμε:

I = {1, 4, 3, 2}

Η αποτελεσματικότητα κόστους του S1 είναι ίση με 5/0.

Η αποτελεσματικότητα κόστους του S2 είναι ίση με 10/1.

Άρα επιλέγεται το S2 και το I γίνεται {1, 2, 3, 4, 5}

I = U, άρα ο αλγόριθμος σταματάει.

Για διαφορετικές συναρτήσεις κόστους μπορεί να δοθεί έμφαση σε διαφορετικά χαρακτηριστικά των υποσυνόλων και επομένως να αλλάξει το σύνολο των υποσυνόλων που επιλέγεται κάθε φορά. Η πιο απλή περίπτωση αποτελεί η συνάρτηση κόστους να είναι ο αριθμός των στοιχείων που περιέχει κάθε φορά το υποσύνολο.

5.2.2 Απόδειξη Λογαριθμικής Προσέγγισης του Αλγορίθμου

Έστω OPT το κόστος της καλύτερης λύσης. Έστω (k-1) τα στοιχεία που καλύπτονται πριν από τυχαία επανάληψη του αλγορίθμου. Το κόστος του k κατά σειρά στοιχείου θα είναι μικρότερο ή ίσο του OPT/(n-k+1). Εφόσον το k κατά σειρά στοιχείο δεν έχει καλυφθεί ακόμα, υπάρχει κάποιο σύνολο Si που δεν έχει επιλεχτεί έως τώρα από τον αλγόριθμο και πρόκειται να επιλεχτεί ως μέρος του OPT. Εφόσον ο άπληστος αλγόριθμος επιλέγει το σύνολο με τη μικρότερη αποτελεσματικότητα κόστους, το κόστος ανά στοιχείο στο επιλεγόμενο σύνολο θα πρέπει να είναι μικρότερο από το OPT διαιρεμένο κατά τον αριθμό των εναπομείναντων στοιχείων. Επομένως, το κόστος του k κατά σειρά στοιχείου είναι μικρότερο ή ίσο του OPT / |U – I|, δηλαδή OPT/(n-k+1).

```
Cost of Greedy Algorithm = Sum of costs of n elements
```

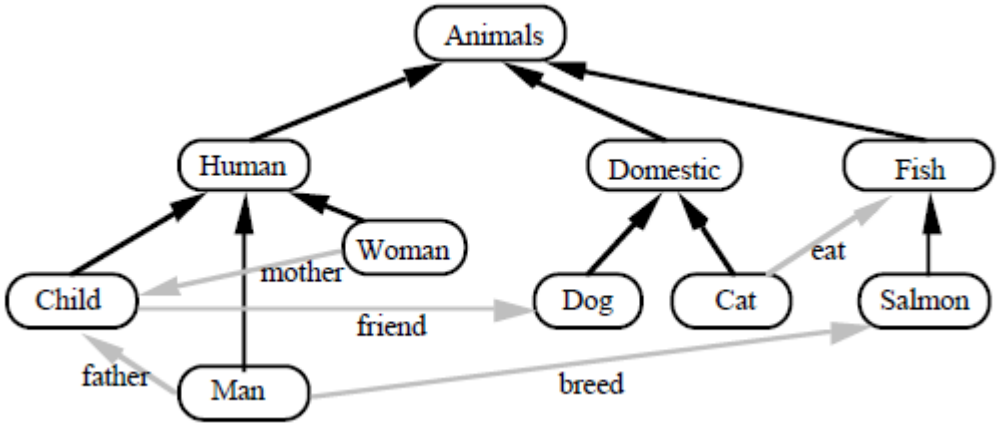
```
[putting k = 1, 2..n in above formula]

  <= (OPT/n + OPT(n-1) + ... + OPT/n)

  <= OPT(1 + 1/2 + ..... 1/n)

[Since 1 + 1/2 + .. 1/n ≈ Log n]

  <= OPT * Logn
```

5.3 Μέτρα ομοιότητας μεταξύ αντικειμένων[10]

Σε αυτό το σημείο, έχοντας έναν άπληστο αλγόριθμο για την εύρεση των καλύτερων δυνατών υποσυνόλων του συνόλου προτιμήσεων, το μόνο που μένει για την εφαρμογή του είναι η εύρεση της κατάλληλης συνάρτησης κόστους. Η συνάρτηση κόστους αυτή όπως έχει διαφανεί στις προηγούμενες σελίδες θα πρέπει να σχετίζεται με την ομοιότητα μεταξύ των στοιχείων του συνόλου κατά τέτοιο τρόπο, ώστε τα σύνολα τα οποία περιέχουν αντικείμενα με όμοια χαρακτηριστικά να έχουν μικρότερο κόστος και συνεπώς να προτιμώνται κατά την εκτέλεση του αλγορίθμου. Παρακάτω παρουσιάζουμε τη βασική μελέτη που έχει γίνει στον τομέα των Τεχνολογιών Γνώσης για τον προσδιορισμό του μέτρου ομοιότητας μεταξύ αντικειμένων. Για το σκοπό αυτό επαναπαρουσιάζουμε κάποιες βασικές έννοιες που αφορούν την Οντολογική Αναπαράσταση Γνώσης.

Βασικά συστατικά στοιχείας μιας οντολογίας είναι οι έννοιες (concepts) και τα αντικείμενα (instances). Τα αντικείμενα αποτελούν απτά ή μη απτά στοιχεία του πεδίου ενδιαφέροντος, όπως πχ μία ταινία ή ένας ηθοποιός. Οι έννοιες αποτελούν κλάσεις ή κατηγορίες στις οποίες ανήκουν ή δεν ανήκουν τα αντικείμενα του πεδίου ενδιαφέροντος. Για παράδειγμα, το αντικείμενο “Star Wars Episode V: The Empire Strikes Back” ανήκει στις κλάσεις “Ταινία Δράσης” και “Ταινία Επιστημονκής Φαντασίας”. Μία Βάση Γνώσης μπορεί να περιγραφεί ως ένας γράφος του οποίου κάθε κόμβος εκφράζει μία έννοια. Η οριζόντια δομή έχει να κάνει με συνδέσμους κληρονομικότητας που υποδεικνύουν την ιεραρχία των κλάσεων, ενώ η κάθετη δομή σχετίζεται με

τις δυαδικές σχέσεις που υπάρχουν μεταξύ δύο εννοιών. Κατά κανόνα ο γράφος αυτός είναι κατευθυνόμενος και ακυκλικός.



Σε αυτά τα πλαίσια, κάθε αντικείμενο ανήκει σε ένα συγκεκριμένο αριθμό εννοιών και κατ' αυτό τον τρόπο συνδέεται με άλλα αντικείμενα του πεδίου ενδιαφέροντος. Η ομοιότητα μεταξύ αυτών των αντικειμένων μπορεί να θεωρηθεί ως “απόσταση” που χωρίζει τα συγκεκριμένα αντικείμενα. Υπό αυτό το πρίσμα, η ομοιότητα έχει μελετηθεί από τις Γνωστικές Επιστήμες και από αυτές θα δανειστούμε τα θεωρητικά εργαλεία για την εισαγωγή ενός μαθηματικού μοντέλου. Ο Amos Nathan Tversky, Μαθηματικός Ψυχολόγος, δίνει έμφαση στις μαθηματικές ιδιότητες που περιγράφει η Ανάλυση Δεδομένων (ελαχιστότητα, συμμετρία, ασυμμετρία) για να ορίσει τον τρόπο με τον οποίο οι άνθρωποι αισθάνονται και αντιμετωπίζουν την έννοια της ομοιότητας. Οι άνθρωποι χρησιμοποιούν τις έννοιες αυτές σε συνδυασμό με τα χαρακτηριστικά των αντικειμένων για να κρίνουν το βαθμό ομοιότητάς τους. Ο Tversky επομένως προτείνει ως μέτρο της ομοιότητας μεταξύ δύο αντικειμένων x και y με σύνολα χαρακτηριστικών A και B αντίστοιχα την εξίσωση:

$$S(x,y) = \frac{f(A \cap B)}{f(A \cup B) + \alpha f(A - B) + \beta f(B - A)}$$

Όπου f μία γραμμική συνάρτηση που δρα πάνω σε σύνολα και α, β θετικοί αριθμοί. Ανάλογα με τον τρόπο με τον οποίο επιλέγουμε τη συνάρτηση f και

τις τιμές των α και β μπορούμε να χρησιμοποιήσουμε διαφορετικά μέτρα ομοιότητας ανάλογα με την περίπτωση. Για παράδειγμα, αν θέλουμε να μετρήσουμε την ομοιότητα μεταξύ δύο αντικειμένων που λαμβάνουν αριθμητικές τιμές, θα μπορούσαμε να χρησιμοποιήσουμε τόσο την έννοια της συμμετρίας, όσο και την έννοια της ασυμμετρίας για να παράγουμε δείκτες ομοιότητας μεταξύ αντικειμένων. Αν για παράδειγμα επιλέξουμε α και β ίσα με 0, τότε για οποιαδήποτε f καταλήγουμε σε ένα συμμετρικό μέτρο ομοιότητας. Αντίθετα, αυξάνοντας την τιμή του α σε σχέση με το β δίνουμε μεγαλύτερο βάρος στο σύνολο A και συνεπώς στο αντικείμενο x, ενώ αντίθετα αυξάνοντας το β, αυξάνουμε τη σημασία που δίνεται στο αντικείμενο y. Υπό αυτό το πρίσμα, θέτοντας τις παραμάετρους α και β ίσες με το 0, προκύπτει το μέτρο της συμμετρικής ομοιότητας το οποίο εκφράζεται ως το πηλίκο:

$$SymSim(x, y) = \frac{f(A \cap B)}{f(A \cup B)}$$

Όπου το σύμβολο SymSim παραπέμπει στο Symmetric Similarity. Το μέτρο που περιγράφει την ασύμμετρη ομοιότητα μπορεί να λάβει πολλές μαθηματικές εκφράσεις ανάλογα με τις τιμές που λαμβάνουν οι παράμετροι. Έτσι, για παράδειγμα, ένα μέτρο ασύμμετρης ομοιότητας που δίνει μεγαλύτερη σημασία στο σύνολο χαρακτηριστικών του x απ' ότι στο σύνολο των χαρακτηριστικών του y, προκύπτει από τον τύπο του Tversky, αν θέσουμε α = 0 και β = -1. Τότε:

$$AsymSim(x, y) = \frac{f(A \cap B)}{f(A)}$$

Όπου το σύμβολο AsymSim παραπέμπει κατ' αντιστοιχεία στο Asymmetric Similarity. Με αυτό τον τρόπο, το μέτρο ασύμμετρης ομοιότητας παρέχει τη δυνατότητα μελέτης του βαθμού συμπερίληψης ανάμεσα στο αντικείμενο x (στοιχείο αναφοράς) και το y (στόχος).

Για την επίλυση του προβλήματος του ανομοιόμορφου συνόλου προτιμήσεων, επιλέχτηκε ένα συμμετρικό μέτρο ομοιότητας, δηλαδή με παραμέτρους α και β ίσες με το 0. Η απόφαση αυτή ήταν εύλογη, εφόσον στη γενική περίπτωση κάθε αντικείμενο που μπορεί να προστεθεί στο σύνολο προτιμήσεων έχει την ίδια σημασία με όλα τα υπόλοιπα. Η συμπερίληψη μεταξύ αντικειμένων δε μας ενδιαφέρει για τους σκοπούς ενός συστήματος συστάσεων. Παρόλα αυτά, θα

πρέπει να σημειωθεί ότι σε ορισμένα συστήματα σύστασης βασισμένα σε γνώση μπορεί να είναι σκόπιμη η επιλογή ενός ασύμμετρου μέτρου ομοιότητας. Η επιλογή της συνάρτησης f έγινε με βάση τη μελέτη των υπαρχόντων μέτρων συμμετρικής ομοιότητας που υπάρχουν στη βιβλιογραφία. Τελικά επιλέχτηκε ως συνάρτηση f ο αριθμός των στοιχείων του συνόλου. Έτσι, η προκύπτουσα συνάρτηση ομοιότητας μεταξύ δύο αντικειμένων x και y που έχουν τα χαρακτηριστικά (ανήκουν στο σύνολο των κλάσεων) A και B έχουν ως βαθμό ομοιότητας:

$$S(x,y) = \frac{|A \cap B|}{|A \cup B|}$$

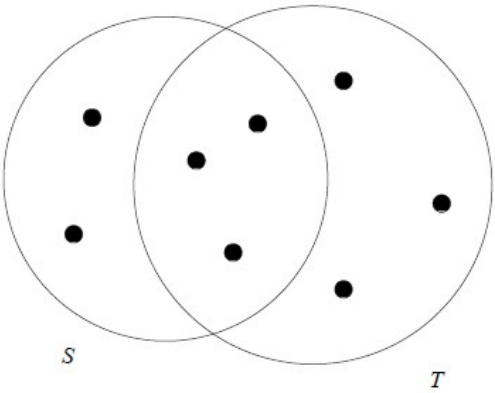


Figure 3.1: Two sets with Jaccard similarity 3/8

Η συγκεκριμένη σχέση παρουσιάζεται γνώριμη από τον τομέα της συνολοθεωρίας, καθώς πρόκειται για την Ομοιότητα Jaccard. Η ταύτιση του συμμετρικού μέτρου ομοιότητας που επιλέχτηκε με την ομοιότητα Jaccard συνδέεται με τρία πολύ σημαντικά πλεονεκτήματα:

- Η ομοιότητα Jaccard παρέχει το απαραίτητο επιστημονικό και ερευνητικό υπόβαθρο για να ενισχύσει το κύρος της συγκεκριμένης επιλογής. Πρόκειται για μέτρο που χρησιμοποιείται εκτενώς σε τομείς όπως η συνολοθεωρία, η γεωλογία και η βιολογία κατά τη μελέτη συνόλων αντικειμένων τα οποία κατέχουν χαρακτηριστικά, είτε σε απόλυτο βαθμό, είτε σε μηδενικό.
- Το μέτρο αυτό μπορεί εύκολα να επεκταθεί μέσω της γενικευμένης ομοιότητας Jaccard κατά τέτοιον τρόπο ώστε να υποστηρίζει συστήματα που μοιάζουν ασαφή (fuzzy-like systems). Στα συστήματα αυτά, τα αντικείμενα μπορεί να ανήκουν στην κλάση κάποιου

χαρακτηριστικού σύμφωνα με ένα βαθμό συμμετοχής, ο οποίος μπορεί να λάβει όλες τις πραγματικές τιμές μεταξύ του 0 και του 1.

- Καλύπτει την ακραία περίπτωση όπου A = B = Ø. Σε αυτή την περίπτωση, S(x, y) = 1 (δύο αντικείμενα χωρίς χαρακτηριστικά είναι όμοια).

Για παράδειγμα, έστω τα αντικείμενα x και y με σύνολα χαρακτηριστικών A = {Action, Adventure, Sci-Fi, Hamill Mark, Ford Harisson, Jones James Earl, Billy Dee Williams} και B = {Action, Adventure, Sci-Fi, Hamill Mark, Ford Harisson, Jones James Earl, Kenneth Colley}. Η ομοιότητα μεταξύ των δύο αντικειμένων είναι:

$$S(x,y) = \frac{|A \cap B|}{|A \cup B|} = \frac{6}{8} = 0.75$$

Αντίθετα, τα αντικείμενα z και w με σύνολα χαρακτηριστικών C = {Documentary, Bush George W, Bush George I, DeNiro Robert, Clinton Bill I) και D = {Romance, Drama, History, Curry Tim, Harris Sam, DiCaprio Leonardo, Winslet Kate} έχουν βαθμό ομοιότητας:

$$S(z,w) = \frac{|A \cap B|}{|A \cup B|} \left(= \frac{0}{0}\right) = 0$$

Όπως αναμενόταν, οι ταινίες “Star Wars Episode V” και “Star Wars Episode VI” έχουν μεγάλο βαθμό ομοιότητας, ενώ οι ταινίες “Fahreneit 9/11” και “Titanic” έχουν μηδενικό βαθμό ομοιότητας.

Σε αυτό το σημείο θα πρέπει να σημειωθεί ότι η αποδοτικότητα του αποτελέσματος της ομοιότητας Jaccard, δηλαδή η δυνατότητα του μέτρου να αποδίδει διαισθητικά σωστά αποτελέσματα, εξαρτάται σε μεγάλο βαθμό από την ποσότητα της γνώσης που έχουμε για το πεδίο ενδιαφέροντος. Για παράδειγμα, εάν στις ταινίες x και y προστεθεί επιπλέον το χαρακτηριστικό “Star Wars”, τότε ο νέος βαθμός ομοιότητας των ταινιών ισούται με 0.78, ενισχύοντας το γεγονός ότι αυτές οι δύο ταινίες μοιάζουν μεταξύ τους.

5.4 Εύρεση της συνάρτησης Cost με βάση την ομοιότητα Jaccard[12]

Όπως έχει αναφερθεί παραπάνω, για την αποδοτική υλοποίηση του άπληστου αλγορίθμου που λύνει το πρόβλημα της κάλυψης συνόλου, απαιτείται η εύρεση κατάλληλης συνάρτησης κόστους για κάθε υποσύνολο. Με βάση την ομοιότητα Jaccard ο υπολογισμός του κόστους ενός υποσυνόλου με δύο μόνο αντικείμενα μπορεί εύκολα να υπολογιστεί ως το συμπλήρωμα ως προς 1 της ομοιότητας μεταξύ των δύο αντικειμένων. Δηλαδή για ένα σύνολο Si που περιέχει τα αντικείμενα x και y:

$$Cost(Si) = 1 - S(x, y)$$

Η επιλογή του συμπληρώματος ως προς 1 γίνεται έτσι ώστε το κόστος να είναι μικρότερο για μεγαλύτερους βαθμούς ομοιότητας.

Η εύρεση της συνάρτησης κόστους γίνεται πιο πολύπλοκη όταν το υποσύνολο του οποίου το κόστος πρέπει να υπολογιστεί περιέχει περισσότερα από δύο αντικείμενα. Σε αυτή την περίπτωση επιλέγουμε να χρησιμοποιήσουμε το μέσο όρο των βαθμών ομοιότητας ανά δύο διορθωμένο κατά έναν παράγοντα n ο οποίος είναι ίσος με τον αριθμό των αντικειμένων του συνόλου προτιμήσεων. Δηλαδή για ένα σύνολο Si που περιέχει τα αντικείμενα i1, i2, ..., in:

$$Cost(Si) = 1 - \frac{n \sum_{i=1, j=1}^{i=n, j=n} S(ki, kj)}{\frac{n!}{2!\,n-2!}}$$

Ο διορθωτικός παράγοντας n επιλέχτηκε ώστε στο σύστημα να προτιμώνται οι μεγαλύτερες ομάδες αντικειμένων. Διαφορετικά, υψηλή ομοιότητα μεταξύ δύο αντικειμένων οδηγεί σε πολύ μικρό κόστος για σύνολα δύο αντικειμένων, μειώνοντας τη σημασία συνόλων τριών ή τεσσάρων αντικειμένων τα οποία θα μπορούσαν να δώσουν καλύτερα αποτελέσματα. Ταυτόχρονα, ο παράγοντας αυτός αυξάνεται γραμμικά με την αύξηση του αριθμού των αντικειμένων, σε αντίθεση με τον παρονομαστή, ο οποίος αυξάνεται παραγοντικά. Σαν αποτέλεσμα, η σημασία του παράγοντα n μειώνεται όσο αυξάνεται ο αριθμός των αντικειμένων και το μέγεθος των συνόλων που επιλέγονται κινείται σχεδόν πάντα σε μονοψήφια πλαίσια.

Για παράδειγμα, το σύνολο των ταινιών x, y, w με σύνολα χαρακτηριστικών A = {Action, Adventure, Sci-Fi, Hamill Mark, Ford Harisson, Jones James Earl, Billy Dee Williams}, B = {Action, Adventure, Sci-Fi, Hamill Mark, Ford Harisson, Jones James Earl, Kenneth Colley} και C = {Romance, Drama, History, Curry Tim, Harris Sam, DiCaprio Leonardo, Winslet Kate} αντίστοιχα έχει κόστος:

$$Cost(Si) = 1 - \frac{n \sum_{i=1,\, j=1}^{i=n,\, j=n} S(ki, kj)}{\frac{n!}{2!\, n-2!}} = 1 - \frac{3(0.75 + 0 + 0)}{3} = 0.25$$

Έχοντας καλύψει τις περιπτώσεις των συνόλων με δύο και παραπάνω αντικείμενα απομένει η εύρεση της τιμής του κόστους για τα μονοσύνολα, τα σύνολα δηλαδή που περιέχουν μόνο ένα αντικείμενο. Η επιλογή του συγκεκριμένου κλάδου της συνάρτησης κόστους είναι ιδιαίτερα σημαντική, καθώς η τιμή αυτή αποτελεί ουσιαστικά την τιμή κατώφλι κάτω από την οποία ένα σύνολο που αποτελείται από δύο ή περισσότερες ταινίες μπορεί να θεωρηθεί αρκετά ομοιογενές για να εκτιμηθεί ως υποψήφιο υποσύνολο του συνόλου προτιμήσεων. Ιδανικά, το κόστος ενός συνόλου με ένα αντικείμενο θα είναι μία σταθερά a, η οποία θα προκύπτει μετά από την υλοποίηση και κυκλοφορία δοκιμαστικής έκδοσης του συστήματος. Στα πλαίσια, ωστόσο της παρούσης εργασίας και για τη λήψη αποτελεσμάτων που παρουσιάζονται σε επόμενο κεφάλαιο, η σταθερά επιλέχτηκε να είναι ίση με 0, τιμή που προέκυψε από εμπειρική παρατήρηση των αποτελεσμάτων του συστήματος.

Συγκεντρώνοντας όλα τα παραπάνω στοιχεία καταλήγουμε στην εξής διακλαδιζόμενη συνάρτηση κόστους:

$$Cost(Si) = 0 \text{ , αν } |Si| = 1$$

$$Cost(Si) = 1 - \frac{n \sum_{i=1,\, j=1}^{i=n,\, j=n} S(ki, kj)}{\frac{n!}{2!\, n-2!}} \text{ , αλλιώς}$$

5.5 Πλήρης Αλγόριθμος Επίλυσης του Προβλήματος Κάλυψης Συνόλου με Χρήση Βαθμού Ομοιότητας

5 Επέκταση του Συστήματος για Πολύπλοκα Σύνολα Προτιμήσεων

Έχοντας την παραπάνω συνάρτηση κόστους μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο που περγράφηκε παραπάνω για να διασπάσουμε το σύνολο προτιμήσεων στα κατάλληλα ομοιογενή υποσύνολα. Όπως έχει αναφερθεί, στόχος της διάσπασης αποτελεί η δημιουργία υποσυνόλων των οποίων τα αντικείμενα θα έχουν υψηλό βαθμό ομοιότητας έτσι ώστε η διαδικασία λήψης συστάσεων να μπορεί να πραγματοποιείται αποδοτικότερα. Τα υποσύνολα του συνόλου προτιμήσεων που προκύπτουν από τη διάσπαση αντιμετωπίζονται με τη σειρά τους ως ξεχωριστά υποσύνολα τα οποία δίνονται ως είσοδος στο σύστημα συστάσεων. Η ένωση των συστάσεων που δίνονται ως έξοδοι αποτελεί και τις τελικές συστάσεις που παρουσιάζονται στο χρήστη.

Ο αλγόριθμος που εσωκλείει όλα τα παραπάνω είναι ο εξής:

Δεδομένα:

1. Ένα ανομοιογενές σύνολο προτιμήσεων U με βάση το οποίο θα πραγματοποιηθούν οι συστάσεις
2. Η συνάρτηση κόστους Cost που δρα πάνω σε σύνολα αντικειμένων

$$0, \text{ αν } |Si| = 1$$

$$, \text{ αλλιώς}$$

Αλγόριθμος:

1. Υπολόγισε το δυναμοσύνολο PS του συνόλου U
2. Αρχικοποίησε το σύνολο S των επιλεγμένων υποσυνόλων στο κενό σύνολο Ø
3. Αρχικοποίηση το σύνολο I των αντικειμένων που έχουν καλυφθεί στο κενό σύνολο Ø
4. Όσο το σύνολο I δεν ταυτίζεται με το σύνολο U επανάλαβε
    1. Για κάθε σύνολο PSi του συνόλου PS υπολόγισε το k(PSi) = Cost(PSi) / |PSi – I|
    2. Πρόσθεσε τα στοιχεία του PSi με το μικρότερο k στο I

3. Αφαίρεσε το σύνολο PSi από το PS
4. Πρόσθεσε το Si στο S
5. Για κάθε σύνολο Si στο S, υπολόγισε τις αντίστοιχες συστάσεις και εμφάνισέ τις

Για παράδειγμα, έστω και πάλι το σύνολο των ταινιών x, y, w με σύνολα χαρακτηριστικών A = {Action, Adventure, Sci-Fi, Hamill Mark, Ford Harisson, Jones James Earl, Billy Dee Williams}, B = {Action, Adventure, Sci-Fi, Hamill Mark, Ford Harisson, Jones James Earl, Kenneth Colley} και C = {Romance, Drama, History, Curry Tim, Harris Sam, DiCaprio Leonardo, Winslet Kate}. Τότε ο αλγόριθμος εκτελείται ως εξής:

- Powerset = {[x], [y], [w], [x, y], [x, w], [y, w], [x, y, w]}
- S = { }
- I = { }
- Πρώτη επανάληψη:
  - k{[x]} = 0
  - k{[y]} = 0
  - k{[w]} = 0
  - k{[x, y]} = -0.25
  - k{[x, w]} = 0.5
  - k{[y, w]} = 0.5
  - k{[x, y, w]} = 0.125
  - Επιλέγεται το σύνολο [x, y]
  - S = {[x, y]}
  - I = {x, y}
- Δεύτερη επανάληψη

- k{[x]} = 1
- k{[y]} = 1

k{[w]} = 0

- k{[x, w]} = 0.75
- k{[y, w]} = 0.75
- k{[x, y, w]} = 0.25
- Επιλέγεται το σύνολο [w]
- S = {[x, y], [w]}
- I = {x, y, w}

- Βρες συστάσεις για το σύνολο [x, y] και εμφάνισέ τις
- Βρες συστάσεις για το σύνολο [w] και εμφάνισέ τις

Ενδεικτικά αποτελέσματα του αλγορίθμου με βάση την κινηματογραφική οντολογία παρουσίαζονται στον παρακάτω πίνακα:

Σύνολο Προτιμήσεων	Επιλεγμένα υποσύνολα Συνόλου Προτιμήσεων
[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi]	[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi]
[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi, Titanic]	[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi], [Titanic]
[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi, Scream 2, Scream3]	[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi], [Scream 2, Scream 3]
[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi, Scream 2, Scream3, Titanic]	[Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi] [Scream 2, Scream 3], [Titanic]
[Star Wars Episode V: The Empire	[Star Wars Episode V: The Empire

Strikes Back, Star Wars Episode VI: Return of the Jedi, Scream 2, Scream3, Mulan, Shrek 2]	Strikes Back, Star Wars Episode VI: Return of the Jedi] [Scream 2, Scream 3], [Star Wars Episode V: The Empire Strikes Back, Star Wars Episode VI: Return of the Jedi, Shrek 2], [Mulan]

Όπως γίνεται αμέσως φανερό από τον παραπάνω πίνακα, η διάσπαση σε υποσύνολα επιφέρει τα εμπειρικώς αναμενόμενα αποτελέσματα σχεδόν στο σύνολο των περιπτώσεων. Η μόνη δυσλειτουργία που παρατηρείται σε αυτά τα ενδεικτικά παραδείγματα σχετίζεται με τη λανθασμένη κατηγοριοποίηση της ταινίας Shrek 2 μαζί με τις ταινίες Star Wars. Ιδανικά, θα θέλαμε η ταινία Shrek 2 να βρίσκεται μαζί με την ταινία Mulan. Το πρόβλημα προκύπτει από το γεγονός ότι η ταινία Shrek 2 ανήκει στις κατηγορίες Action και Adventure, τις οποίες μοιράζεται με τις ταινίες Star Wars που έχουμε συμπεριλάβει στο σύνολο προτιμήσεων του χρήστη.

6 Ζητήματα Σχεδίασης της Διεπαφής και Μελλοντικές Βελτιώσεις

Όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο, οι τρεις βασικοί άξονες που έχουν ενδιαφέρον κατά τη σχεδίαση μίας εφαρμογής παροχής συστάσεων βασισμένων σε γνώση είναι:

- ο τρόπος με τον οποίο ο χρήστης εισάγει τις απαιτήσεις του στο σύστημα,
- η βασική λογική με την οποία γίνονται οι συστάσεις (ο αλγόριθμος επιλογής των συστάσεων) και
- τα εργαλεία που παρέχονται στο χρήστη για να εφαρμόσει την κριτική μέθοδο στα αποτελέσματα.

Τα συστήματα σύστασης βασισμένα σε γνώση χωρίζονται σε δύο μεγάλες κατηγορίες με βάση τον τρόπο με τον οποίο ο χρήστης εισάγει τις απαιτήσεις του στο σύστημα:

1. Συστήματα βασισμένα σε περιορισμούς. Στα συστήματα σύστασης βασισμένα σε περιορισμούς, οι χρήστες τυπικά επιλέγουν απαιτήσεις ή περιορισμούς (πχ συγκεκριμένες τιμές) για τα χαρακτηριστικά των αντικειμένων. Επιπλέον, κανόνες που αναφέρονται στο συγκεκριμένο πεδίο χρησιμοποιούνται για να συνδέσουν απαιτήσεις των χρηστών με χαρακτηριστικά των αντικειμένων (για παράδειγμα η απαίτηση του χρήστη για την εύρεση αυτοκινήτων που έχουν αυτόματα πλοηγό αποκλείει αυτομάτως όλα τα αυτοκίνητα που έχουν κατασκευαστεί πριν το 2005). Ανάλογα με τον αριθμό και τον τύπο των επιστρεφόμενων αποτελεσμάτων, ο χρήστης μπορεί να έχει την ευκαιρία να αλλάξει τις αρχικές απαιτήσεις. Για παράδειγμα, μπορεί να χαλαρώσει κάποιους περιορισμούς, εάν επιστραφούν πολύ λίγες προτάσεις ή να προσθέσει επιπλέον εάν συμβαίνει το αντίθετο. Αυτή η διαδικασία αναζήτησης επαναλαμβάνεται εώς ότου ο χρήστης φτάσει στα επιθυμητά αποτελέσματα.
2. Συστήματα σύστασης βασισμένα σε περιπτώσεις: Στα συστήματα σύστασης βασισμένα σε περιπτώσεις, ορισμένες περιπτώσεις

δηλώνονται από το χρήστη ως στόχοι. Μέτρα ομοιότητας ορίζονται ώστε να επιλεχθούν συγκεκριμένα χαρακτηριστικά και να συσταθούν τα κατάλληλα αντικείμενα. Τα μέτρα αυτά συνήθως είναι προσεκτικά επιλεγμένα με βάση το πεδίο ενδιαφέροντος. Επομένως, πρόκειται για διαφορετικά μέτρα ανάλογα με τη φύση του συστήματος και τους στόχους λειτουργίας. Τα επιστρεφόμενα αποτελέσματα χρησιμοποιούνται συνήθως ως νέοι στόχοι με κάποιες διαδραστικές διορθώσεις από το χρήστη για την παροχή των τελικών συστάσεων. Για παράδειγμα, όταν ένας χρήστης δει κάποιο επιστρεφόμενο αποτέλεσμα το οποίο σχεδόν ταιριάζει με αυτό που θέλει, μπορεί να επαναλάβει την αναζήτηση με αυτό το αντικείμενο ως στόχο και ενδεχομένως με κάποιες αλλαγές στα χαρακτηριστικά του κατά τις προτιμήσεις του. Αυτή η διαδραστική διαδικασία μπορεί να οδηγήσει πιο αποτελεσματικά το χρήστη στις επιθυμητές συστάσεις.

Όπως έχει ήδη διαφανεί από την περιγραφή της βασικής λογικής με την οποία γίνονται οι συστάσεις, το σύστημα που δημιουργήθηκε είναι ένα σύστημα σύστασης βασισμένο σε περιπτώσεις. Η εφαρμογή που δημιουργήθηκε για να αναδείξει τη λειτουργία του συστήματος είναι μία διαδικτυακή εφαρμογή για την οποία αναπτύχθηκαν όλα τα components, από το backend και το data modeling έως το frontend και τα γραφικά της διεπαφής. Το backend κομμάτι της διαδικτυακής εφαρμογής υλοποιήθηκε σε Java, ενώ υπεύθυνος για τη διαχείριση των αιτήσεων (requests) ήταν ένας Java Servlet. Το frontend κομμάτι υλοποιήθηκε σε HTML, CSS, Javascript και Java χάρη στη χρήστη της τεχνολογίας των Java Server Pages.

Η αρχική σελίδα της εφαρμογής περιέχει μία μπάρα αναζήτησης μαζί με το λογότυπο της εφαρμογής και ένα κουμπί για την εκκίνηση της διαδικασίας των συστάσεων. Ο χρήστης εισάγει ονόματα ταινιών που τον ενδιαφέρουν μέσω της μπάρας αναζήτησης και με το πάτημα του πλήκτρου Enter ή μέσω του κλικ στο αντίστοιχο κουμπί, η διαδικασία των συστάσεων ξεκινάει. Δίνεται η δυνατότητα εισαγωγής μίας ή περισσότερων ταινιών αρκεί στην περίπτωση των πολλών ταινιών αυτές να χωρίζονται με κόμμα. Με το πάτημα του Enter ή με το κλικ γίνεται postback στον server, ο οποίος λαμβάνει το περιεχόμενο της μπάρας ως συμβολοσειρά και μετά την κατάλληλη επεξεργασία τροφοδοτεί το περιεχόμενο στη μηχανή συστάσεων. Τα

αποτελέσματα αποθηκεύονται σε μία λίστα και προωθούνται προς αξιοποίηση σε μία Java Server Page.

Η παρουσίαση των αποτελεσμάτων αποτελεί πιο πολύπλοκη υπόθεση από την απλή παράθεση των τίτλων των ταινιών που επιστρέφει η μηχανή συστάσεων. Προκειμένου να ενισχυθεί η εμπειρία του χρήστη και να διευκολυνθεί η μεταφορά της πληροφορίας, απαιτείται μία πιο παραστατική μορφή παρουσίασης. Γι' αυτό το λόγο επιλέχτηκε η δημιουργία ενός πλέγματος (grid) των αφισών των ταινιών κάθε μία από τις οποίες συνοδεύεται από την αντίστοιχη βαθμολογία που έχει λάβει η ταινία στο tMDB. Προκειμένου να επιτευχθεί αυτός ο στόχος γίνεται χρήση της Διεπαφής Προγραμματισμού Εφαρμογών[13] (API) που ενσωματώνει το tMDB η οποία με τις κατάλληλες κλήσεις επιστρέφει σχετικά στοιχεία από τη βάση δεδομένων που συντηρείται από το tMDB. Ένα διαδικτυακό API είναι ένα API που έχει σχεδιαστεί για να χρησιμοποιηθεί είτε από έναν διακτυακό server, είτε από ένα φυλλομετρητή πλοήγησης. Πρόκειται για μία έννοια στο δικτυακό προγραμματισμό, η οποία συνήθως χρησιμοποιείται κατά την εκτέλεση κώδικα client-based κώδικα για την απόκτηση πρόσβασης σε δημόσια διαθέσιμα δεδομένα και υπηρεσίες. Τέτοιου είδους APIs παρέχουν χιλιάδες διαδικτυακές υπηρεσίες και βάσεις δεδομένων με σκοπό την απρόσκοπτη διακίνηση των δημόσιων πληροφοριών που επεξεργάζονται και τη διευκόλυνση της ανάπτυξης διαδικτυακών εφαρμογών. Τέτοια site συμπεριλαμβάνουν το Yahoo, το Google, το Facebook, το Instagram, το Ebay, το IMDB και άλλα. Ο τρόπος με τον οποίο δουλεύουν τα APIs είναι μέσω http requests τα οποία γίνονται με τη χρήση URLs που περιγράφουν συγκεκριμένες υπηρεσίες που εκτελούνται με βάση κάποια ιστοσελίδα ή διαδικτυακή εφαρμογή.

Σε αυτά τα πλαίσια, το tMDB παρέχει τρεις βασικές υπηρεσίες μέσω του API του, τις Find, Discover και Search. Και στις τρεις περιπτώσεις η απάντηση που επιστρέφει στο σύστημα που κάνει το http request είναι μία JSON. Η επεξεργασία των δεδομένων της JSON σελίδας μπορεί να γίνει είτε με απλή επεξεργασία του κώδικα JSON, είτε με ειδικά εργαλεία της γλώσσας Java όπως η βιβλιοθήκη org.json. Πιο αναλυτικά οι τρεις υπηρεσίες του tMDB είναι οι εξής:

- Μέσω της υπηρεσίας Find, ο χρήστης μπορεί, χρησιμοποιώντας το ID της ταινίας που τον ενδιαφέρει να κάνει αναζήτηση για τα στοιχεία που

είναι διαθέσιμα στο tMDB για τη συγκεκριμένη ταινία. Πιο συγκεκριμένα, η μορφή του URL που χρησιμοποιείται με το http request είναι:

https://api.themoviedb.org/3/find/{external_id}?
api_key=<<api_key>>&language=en-US&external_source=imdb_id

όπου external ID το ID της συγκεκριμένης ταινίας από το iMDB και api key, ένα ειδικό κλειδί μοναδικό για κάθε χρήστη του API του tMDB.

- Μέσω της υπηρεσίας Discover, ο χρήστης μπορεί να αναζητήσει ταινίες με βάση τα χαρακτηριστικά τους, όπως βαθμολογίες, είδη, ηθοποιούς και άλλα.
- Μέσω της υπηρεσίας Search, ο χρήστης μπορεί χρησιμοποιώντας το όνομα μίας ταινίας να αναζητήσει στη βάση δεδομένων του tMDB όλες τις ταινίες που έχουν το συγκεκριμένο όνομα. Στη συντριπτική πλειονότητα των περιπτώσεων το αίτημα αυτό επιστρέφει ένα JSON αρχείο με στοιχεία για μία μόνο ταινία, αλλά αυτό δεν ισχύει απαραίτητα. Για παράδειγμα, σε περιπτώσεις reboot ταινιών με το ίδιο ακριβώς όνομα, όπως για παράδειγμα στην περίπτωση του Fantastic 4, επιστρέφονται δύο ταινίες. Το ίδιο συμβαίνει και με ταινίες από διαφορετικά στούντιο και παραγωγούς που τυγχάνει να έχουν το ίδιο όνομα. Για παράδειγμα, ο Ηρακλής της Disney (1997) μοιράζεται το ίδιο όνομα με την ταινία “Ηρακλής” του 2014 που έχει ως πρωταγωνιστή τον Dwayne Johnson. Παρόλα αυτά, αυτές οι περιπτώσεις είναι μεμονωμένες και συνήθως, αν επιστραφούν παραπάνω από δύο αποτελέσματα, η πρώτη ταινία είναι αυτή που περιέχει τα στοιχεία που ενδιαφέρουν το χρήστη.

Όπως γίνεται αντιληπτό από την περιγραφή των υπηρεσιών, αυτή που χρησιμοποιήθηκε για την εκμαίευση των αφισών είναι η υπηρεσία Search. Κι αυτό γιατί μετά το τέλος της λειτουργίας της μηχανής συστάσεων, η έξοδος που δίνεται στην εφαρμογή είναι μία λίστα συμβολοσειρών με ονόματα ταινιών. Έτσι, για παράδειγμα, αν αναζητήσουμε με βάση το όνομα ταινίας “Star Wars Episode V The Empire Strikes Back” λαμβάνουμε το εξής αποτέλεσμα:

```
// 20170710112525
// https://api.themoviedb.org/3/search/movie?api_key=cc4b67c52acb514bdf4931f7cedfd12b&query=Star-Wars-Episode-V-The-Empire-Strikes-Back

{
  "page": 1,
  "total_results": 1,
  "total_pages": 1,
  "results": [
    {
      "vote_count": 5306,
      "id": 1891,
      "video": false,
      "vote_average": 8.2,
      "title": "The Empire Strikes Back",
      "popularity": 4.826687,
      "poster_path": "/6u1fYtxG5eqjhtCPDx04pJphQRW.jpg",
      "original_language": "en",
      "original_title": "The Empire Strikes Back",
      "genre_ids": [
        12,
        28,
        878
      ],
      "backdrop_path": "/amYkOxCwHiVTFKendcIW0rSrRlU.jpg",
      "adult": false,
      "overview": "The epic saga continues as Luke Skywalker, in hopes of defeating the evil Galactic Empire, learns the ways of the Jedi from aging master Yoda. But Darth Vader is more determined than ever to capture Luke. Meanwhile, rebel leader Princess Leia, cocky Han Solo, Chewbacca, and droids C-3PO and R2-D2 are thrown into various stages of capture, betrayal and despair.",
      "release_date": "1980-05-17"
    }
  ]
}
```


Το αποτέλεσμα που επιστρέφεται είναι ένα, ενώ τα στοιχεία που μας ενδιαφέρουν είναι το vote_average και το poster_path. Το vote_average είναι ο μέσος όρος των βαθμολογιών που έχει λάβει η ταινία από τους χρήστες του tMDB, ενώ το poster_path είναι μέρος του URI που προσδιορίζει την αφίσα της ταινίας. Πιο συγκεκριμένα, για να μπορέσει να εμφανιστεί η αφίσα απαιτείται η χρήση του URL:

http://image.tmdb.org/t/p/w500/6u1fYtxG5eqjhtCPDx04pJphQRW.jpg

Αν επισκεφθούμε τη συγκεκριμένη διεύθυνση θα λάβουμε την αφίσα της ταινίας, η οποία με απλή HTML και CSS μπορεί να ενσωματωθεί στο πλέγμα που παρουσιάζεται στο χρήστη. Το ίδιο πλέγμα χρησιμοποιείται διαδραστικά για την εφαρμογή της κριτικής μεθόδου. Ο χρήστης μπορεί να κάνει κλικ σε οποιαδήποτε ταινία που παρουσιάστηκε ως αποτέλεσμα για να επανεκκινήσει τη διαδικασία των συστάσεων με σύνολο προτιμήσεων το μονοσύνολο που περιέχει την αντίστοιχη ταινία. Σε αυτά τα πλαίσια, πρόκειται για ένα σύστημα ολικής κριτικής αποτελεσμάτων, ένα σύστημα δηλαδή σύμφωνα με το οποίο όλα τα χαρακτηριστικά του αντικειμένου με βάση το οποίο έγινε η αρχική αναζήτηση αντικαθίστανται από τα χαρακτηριστικά του νέου επιλεγμένου αντικειμένου. Ο σκοπός της εφαρμογής της συγκεκριμένης κριτικής μεθόδου είναι η εξερεύνηση των ταινιών από το χρήστη και η ανακάλυψη επιπλέον ταινιών που μπορεί να τον ενδιαφέρουν. Εξάλλου, όπως έχει αναφερθεί είναι δύσκολο για οποιονδήποτε χρήστη να δηλώσει με μία προσπάθεια επακριβώς το σύνολο του φάσματος των ενδιαφερόντων του.

Η αποτελεσματικότητα του συνόλου των μεθόδων που χρησιμοποιήθηκαν παρουσιάζονται στο επόμενο και τελευταίο κεφάλαιο της παρούσης εργασίας. Παρόλα αυτά, θεωρείται σκόπιμη η αναφορά ορισμένων ιδεών που μπορούν δυνητικά να βελτιώσουν την εμπειρία του χρήστη και κατ' επέκταση την αποτελεσματικότητα της μηχανής συστάσεων.

Αρχικά, το πρώτο ζήτημα που εντοπίστηκε ήταν το γεγονός ότι σε αυτή τη φάση της ανάπτυξης της εφαρμογής, προκειμένου ο χρήστης να μπορέσει να βρει προτάσεις με βάση μία ταινία που τον ενδιαφέρει οφείλει να εισάγει επακριβώς το όνομα της ταινίας στη μπάρα αναζήτησης. Διαφορετικά, το σύστημα δεν αναγνωρίζει την ταινία και επιστρέφει ως σύνολο συστάσεων το κενό σύνολο. Έτσι, αν κάποιος χρήστης εισάγει “The Matrix”, αντί για “Matrix” στο σύστημα δε λαμβάνει τις αντίστοιχες συστάσεις. Προκειμένου να επιλυθεί αυτό το πρόβλημα, εκτιμήθηκε αρχικά η χρήση ενός μέτρου σύγκρισης μεταξύ συμβολοσειρών, που ονομάζεται απόσταση Levenstein. Πρόκειται για μετρική που μπορεί να υλοποιηθεί με τη μορφή συνάρτησης που δέχεται ως ορίσματα δύο συμβολοσειρές και υπολογίζει τον ελάχιστο αριθμό αλλαγών (προσθήκες, αφαιρέσεις, εναλλαγές χαρακτήρων) για τη μετατροπή από της μίας από αυτές τις συμβολοσειρές στην άλλη. Χρησιμοποιώντας τη μετρική αυτή μπορούμε να συγκρίνουμε τη συμβολοσειρά που δίνει ο χρήστης με όλα τα ονόματα ταινιών που είναι διαθέσιμες στη βάση γνώσης και συνεπώς να βρούμε τον πιο κοντινό τίτλο που ταιριάζει στον τίτλο που εισήγαγε ο χρήστης.

Επιπλέον, θα ήταν χρήσιμο να αναλυθεί αλγόριθμος για την προσθήκη νέων αντικειμένων στο γράφο χωρίς να απαιτείται η ανακατασκευή του. Μία ιδέα γύρω από την οποία θα μπορούσε να στηριχθεί η ανάλυση θα ήταν η εύρεση όλων των κόμβων των οποίων η ετικέτα αποτελεί υποσύνολο του συνόλου των εννοιών στις οποίες ανήκει το νέο αντικείμενο και η προσθήκη του αντικειμένου αυτού σε αυτούς τους κόμβους. Ακόμα πιο χρήσιμη θα ήταν η αυτοματοποίηση της διαδικασίας με έλεγχο για την εμφάνιση καινούριων ταινιών στο IMDB και το tMDB. Όπως έχει αναφερθεί, η καινοτομία αποτελεί βασικό στόχο των συστημάτων σύστασης και επομένως η συνεχής ενημέρωση της βάσης καθίσταται επιτακτική ανάγκη για την πετυχημένη λειτουργία του συστήματος και την εξασφάλιση της αντοχής του στο χρόνο.

Τέλος, θα μπορούσε να ήταν χρήσιμη η εισαγωγή cookies στην εφαρμογή, ώστε να μπορεί να χρησιμοποιεί ο χρήστης πιο αποτελεσματικά τις προηγούμενες αναζητήσεις του – χωρίς φυσικά αυτές να χρησιμοποιούνται στη διαδικασία των συστάσεων.

7. Σύγκριση Αποτελεσμάτων Συστήματος με τη Βιβλιοθήκη Συστημάτων Σύστασης Apache Mahout

Ως αυτό το σημείο έχει περιγραφεί ο τρόπος με τον οποίο λειτουργεί το σύστημα συστάσεων που σχεδιάστηκε, μαζί με κάποιες αποφάσεις που πάρθηκαν για τη βελτίωση της λειτουργίας του. Όπως έχει γίνει φανερό μέσω παραδειγμάτων, το σύστημα επιδιώκει – και καταφέρνει – να πετύχει τους τέσσερις βασικούς στόχους που έχουν όλα τα συστήματα σύστασης, δηλαδή τη σχετικότητα, την καινοτομία, την εναλλακτικότητα και την ποικιλομορφία. Παρόλα αυτά, θεωρείται σκόπιμη η σύγκριση του συστήματος που κατασκευάστηκε στα πλαίσια της παρούσας εργασίας με ένα ήδη υπάρχον πετυχημένο σύστημα προκειμένου να τεθεί υπό καλύτερη προοπτική το ποσοστό επιτυχίας της λειτουργίας του. Σε αυτά τα πλαίσια, επιλέχτηκε το σύστημα Apache Mahout, το οποίο παρέχει εκτεταμένο κατάλογο εργαλείων για χρήση σε συστήματα σύστασης. Παρακάτω γίνεται σύντομη αναφορά στα εργαλεία σύστασης του Mahout και παρουσιάζεται μία στατιστική συγκριτική μελέτη του δικού μας συστήματος σε σχέση με το Mahout[14].

Μία μηχανή σύστασης συνεργατικού φιλτραρίσματος βασισμένη στο Mahout λαμβάνει τις προτιμήσεις των χρηστών και επιστρέφει εκτιμώμενες προτιμήσεις για άλλα αντικείμενα. Το Mahout παρέχει μεγάλη ποικιλία εργαλείων για την κατασκευή ιδιόμορφων συστημάτων σύστασης από μία ευρεία συλλογή αλγορίθμων. Το Mahout έχει σχεδιαστεί για να παρέχει στους προγραμματιστές μεγάλη ελαστικότητα και υψηλή ακρίβεια συστάσεων. Τα βασικά συστατικά στοιχεία των συστημάτων σύστασης που παράγονται από το Mahout έχουν τις εξής βασικές αφαιρετικές μονάδες:

- DataModel
- UserSimilarity
- ItemSimilarity
- UserNeighborhood
- Recommender

Επιλέγοντας διαφορετικούς συνδυασμούς από τις διαθέσιμες επιλογές για κάθε μία από αυτές τις πέντε μονάδες προκύπτουν διαφορετικά συστήματα σύστασης. Το DataModel δηλώνει τον τρόπο με τον οποίο εισάγονται τα δεδομένα που υπάρχουν ήδη για τους χρήστες στο σύστημα, πχ plain txt document ή είσοδος από κάποια βάση δεδομένων. Το UserSimilarity ορίζει την έννοια της ομοιότητας μεταξύ των χρηστών. Πρόκειται για το βασικό στοιχείο στη δημιουργία συστημάτων σύστασης συνεργατικού φιλτραρίσματος μεταξύ χρηστών. Το UserNeighborhood ορίζει τα όρια μέσα στα οποία το σύστημα θα επιδιώξει να βρει παρόμοιους χρήστες και με βάση αυτούς να πραγματοποιήσει τις συστάσεις. Για παράδειγμα, οι δέκα κοντινότεροι χρήστες. Ανάλογο χαρακτήρα με το UserSimilarity έχει το ItemSimilarity, μόνο που σε αυτή την περίπτωση ορίζεται η ομοιότητα μεταξύ αντικειμένων. Το ItemSimilarity χρησιμοποιείται για τη δημιουργία συστημάτων σύστασης συνεργατικού φιλτραρίσματος μεταξύ αντικειμένων. Τέλος, η μονάδα Recommender ορίζει το βασικό αλγόριθμο με τον οποίο εξάγονται οι συστάσεις.

Για τις ανάγκες της σύγκρισης επιλέχτηκε η δημιουργία ενός συστήματος σύστασης συνεργατικού φιτλραρίσματος μεταξύ αντικειμένων, καθώς αποτελεί μοντέλο πιο κοντινό στα συστήματα σύστασης βασισμένων σε γνώση απ' ότι ένα αντίστοιχο σύστημα φιλτραρίσματος μεταξύ χρηστών. Στο DataModel εισήχθησαν με τη μορφή βάσης δεδομένων όλα τα στοιχεία για τις βαθμολογίες από το ML Small Dataset, ενώ το ItemSimilarity που επιλέχτηκε ήταν το PearsonCorrelationSimilarity, το οποίο χρησιμοποιεί τις ομώνυμες μαθηματικές σχέσεις για να υπολογίσει την ομοιότητα μεταξύ των αντικειμένων. Η λειτουργία του συστήματος σύστασης που επιλέχτηκε για κάθε ταινία κατέτασσε τις υπόλοιπες ταινίες με βάση την ομοιότητά τους με τη συγκεκριμένη ταινία και έδινε έναν αριθμό από το 0 έως το 1 που εξέφραζε το βαθμό ομοιότητας κάθε ζευγαριού ταινιών. Προτού προβούμε στη στατιστική ανάλυση των αποτελεσμάτων, θα πρέπει να σημειωθεί ότι ο μέσος όρος των βαθμών ομοιότητας των ταινιών που προτείνει το Mahout για κάθε μία ταινία ξεχωριστά είναι μόλις 0.5536, γεγονός το οποίο καταδεικνύει ότι ακόμα και τα δοκιμασμένα συστήματα δεν επιτυγχάνουν πάντα να βρουν πλήρεις ταυτίσεις μεταξύ αντικειμένων, καθώς σε πολλές περιπτώσεις αυτή η ταύτιση δεν υπάρχει καν. Επομένως, με την παρακάτω σύγκριση δεν

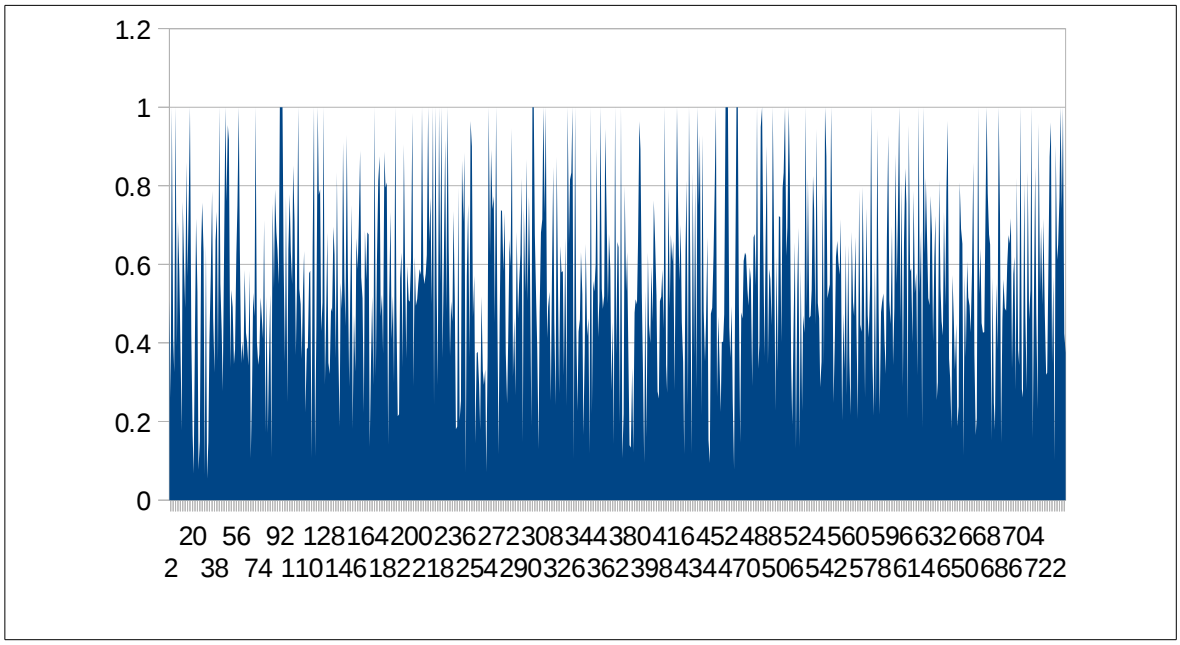
περιμένουμε τέλεια αποτελέσματα της τάξεως του 90 και του 100%, αλλά αποτελέσματα τα οποία ρεαλιστικά είναι καταδεικνύουν μία ομοιομορφία μεταξύ των ταινιών που προτείνονται από το Mahout και ταινιών που προτείνονται από το σύστημα σύστασης βασισμένο σε γνώση που σχεδιάσαμε.

Σε αυτά τα πλαίσια, πάρθηκαν τρεις μετρικές για τη σύγκριση των δύο συστημάτων:

1. Ο αριθμός των κοινών ταινιών που προτείνουν και τα δύο συστήματα.
2. Ο μέσος όρος των βαθμολογιών που δίνει το Mahout σε αυτές τις κοινές ταινίες.
3. Ο μέσος όρος των βαθμολογιών που δίνει το Mahout για όλες τις ταινίες που προτείνονται από το σύστημα βασισμένο σε γνώση κανονικοποιημένες ως προς το μέσο όρο των βαθμών ομοιότητας του Mahout – που είναι ίσος με 0.5536.

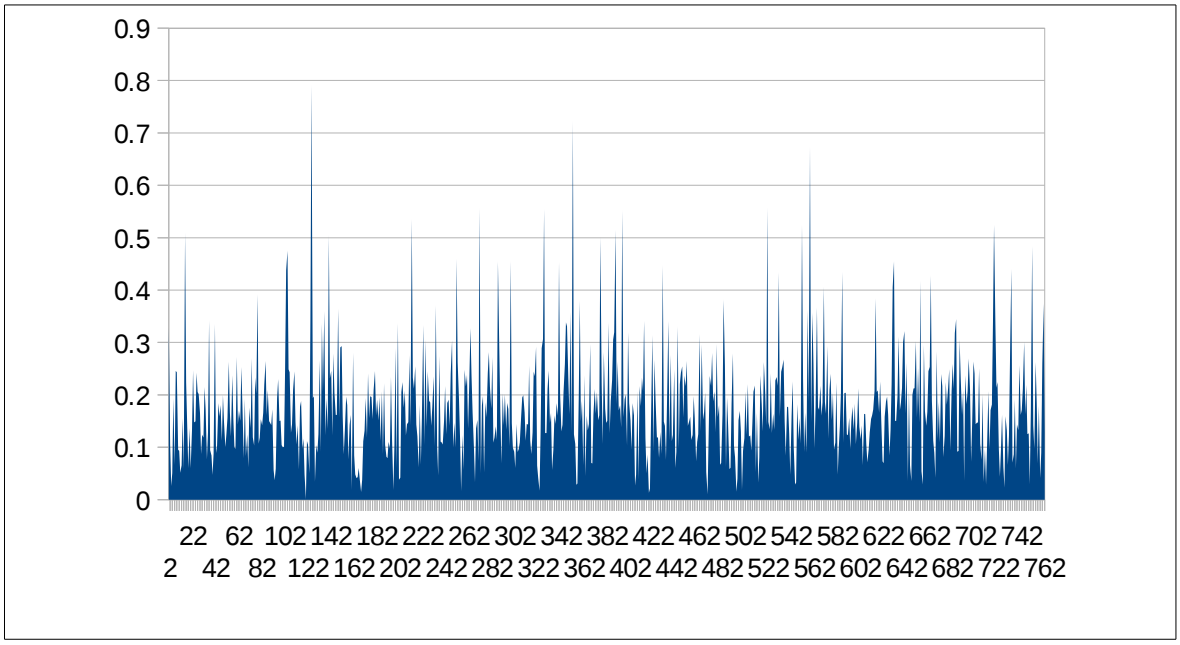
Ως δείγμα επιλέχτηκαν 764 τυχαίες ταινίες από το ML Small Dataset.

Σε αυτά τα πλαίσια, ο μέσος όρος των βαθμολογιών που δίνει το Mahout για τις κοινές ταινίες φαίνεται στο παρακάτω διάγραμμα για τις 764 ταινίες.



Ο μέσος όρος των μέσων όρων είναι τελικά ίσος με 0.5534, αριθμός ο οποίος είναι εξαιρετικά κοντά στο γενικότερο μέσο όρο που έχει το Mahout. Αυτό σημαίνει ότι γενικότερα, οι κοινές ταινίες που προτείνονται και από τα δύο

συστήματα προτείνονται στο Mahout ως πρώτες συστάσεις, top recommendations. Από εκεί και πέρα, ωστόσο, όταν θεωρήσουμε το σύνολο των ταινιών που προτείνει το σύστημά μας σε σχέση με το Mahout, το ποσοστό πέφτει αρκετά, αγγίζοντας κατά μέσο όρο το 0.185:



Το γεγονός αυτό, ωστόσο, είναι εύλογο, αν αναλογιστεί κανείς τα εξής:

1. Το σύστημά μας είναι ένα σύστημα που σκοπό έχει να αντιμετωπίσει το πρόβλημα της ψυχρής εκκίνησης. Επομένως, δεν αξιοποιεί κανένα στοιχείο από προηγούμενη χρήση της εφαρμογής, παρά μόνον τη γνώση που υπάρχει για το πεδίο ενδιαφέροντος.

2. Διαφορετικά συστήματα σύστασης με εξίσου ευρεία αποδοχή συστήνουν τελείως διαφορετικές ταινίες για το ίδιο σύνολο προτιμήσεων του χρήστη. Για παράδειγμα, για το The Force Awakens, εκτός από τις προφανείς συστάσεις (τα υπόλοιπα Star Wars δηλαδή) το IMDB και το tMDB προτείνουν τελείως διαφορετικές ταινίες. Το iMDB κινείται περισσότερο σε μοντέρνες συστάσεις επιλέγοντας ταινίες με έντονο το στοιχείο της φαντασίας. Έτσι, προτείνει το Deadpool, τους Avengers, το Civil War και το Lord of the Rings, ενώ το tMDB παραθέτει ταινίες με βάση τον Action, Adventure και Sci-Fi χαρακτήρα του Force Awakens, προτείνοντας το The Martian, το Spectre, το Hateful Eight και το Ant-Man. Σημαίνει αυτό ότι το ένα σύστημα συστάσεων είναι καλύτερο από το άλλο; Όχι απαραίτητα. Απλά είναι διαφορετική η φιλοσοφία παροχής

των συστάσεων. Στην πραγματικότητα, ο κριτής της αποτελεσματικότητας του συστήματος είναι τελικά ο χρήστης και όχι ένας τυπικός αριθμός που προκύπτει από σύγκριση συμβολοσειρών.

Βιβλιογραφία

1. Brachman, Ronald J., and Hector J. Levesque. *Knowledge Representation and Reasoning*. Amsterdam: Elsevier, 2009

2. Στάμου, Γ., 2015. *Αναπαράσταση οντολογικής γνώσης και συλλογιστική*. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

3. Abiteboul, Serge. *Web Data Management*. Cambridge: Cambridge UP, 2012

4. Harper, F. Maxwell, and Joseph A. Konstan. "The Movielens Datasets: History and Context."*Transactions on Interactive Intelligent Systems*. Association for Computing Machinery (ACM), 15 June 2016

5. Herr, Bruce W., Weimao Ke, Elisha Hardy, and Katy Borner. "Movies and Actors: Mapping the Internet Movie Database." *2007 11th International Conference Information Visualization (IV '07)* (2007): n. pag. Web

6. Aggarwal, Charu C. "Recommender Systems: The Textbook 1st Ed. 2016 Edition."*Recommender Systems: The Textbook: Charu C. Aggarwal “*

7. Antoniou, Grigoris, and Frank Van Harmelen. "Web Ontology Language: OWL." *Handbook on Ontologies* (2009): 91-110. Print

8. Gabriel, Jeff. *Professional .NET Framework*. Birmingham: Wrox, 2001. Print

9. Michalis Giazitzoglou, Alexandros Chortaras, and Giorgos Stamou. “Automatic suggestion of terms for querying data through EL ontologies” (in progress)

10. Gilles Bisson, “Why and How to Define a Similarity Measure for Object Based Representation Systems”, 1995

11. Chandu, Drona Pratap. "Improved Greedy Algorithm for Set Covering Problem." *[1506.04220] Improved Greedy Algorithm for Set Covering Problem*. N.p., 13 June 2015

12. Leskovec, Jurij, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of Massive Datasets*. Cambridge: Cambridge UP, 2015. Print

13. *API Overview - The Movie Database (TMDb)*. N.p., n.d. Web. 20 July 2017

14. Carlos E. Seminario and David C. Wilson, “Case Study Evaluation of Mahout as a Recommender Platform”, 23 September 2012