
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ


# Προσαρμογή και αξιολόγηση μεθόδων εξατομίκευσης αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της


**ΕΥΜΟΡΦΙΑΣ ΜΠΙΛΙΡΗ**



**Επιβλέπων :** Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.



Αθήνα, Δεκέμβριος 2012

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

# Προσαρμογή και αξιολόγηση μεθόδων εξατομίκευσης αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

**ΕΥΜΟΡΦΙΑΣ ΜΠΙΛΙΡΗ**

**Επιβλέπων :** Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Δεκεμβρίου 2012.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Τιμολέων Σελλής | Ιωάννης Βασιλείου | Θοδωρής Δαλαμάγκας |
| Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. | Ερευνητής Β' ΙΠΣΥ/Ε.Κ. "Αθηνά" |

Αθήνα, Δεκέμβριος 2012

.....

**ΕΥΜΟΡΦΙΑ ΜΠΙΛΙΡΗ**

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

# Περίληψη

Η σταδιακή μετεξέλιξη του Παγκόσμιου Ιστού, ενός ιστού εγγράφων, στο Σημασιολογικό Ιστό, έναν ιστό δεδομένων, δημιουργεί την ανάγκη ανάπτυξης μηχανών αναζήτησης σε σημασιολογικά, δομημένα δηλαδή, δεδομένα. Παράλληλα, καθώς ο όγκος της πληροφορίας που προστίθεται καθημερινά στον ιστό αυξάνεται με αλματώδεις ρυθμούς, γίνεται επιτακτική η ανάγκη εφαρμογής μεθόδων εξατομίκευσης από τις νέες αυτές μηχανές αναζήτησης ώστε να προσαρμόζονται στα ενδιαφέροντα κάθε χρήστη.

Στα πλαίσια της παρούσας διπλωματικής εργασίας α)Μελετήθηκαν οι απαιτήσεις και οι περιορισμοί που επιβάλλει το RDF μοντέλο στην αναζήτηση με λέξεις κλειδιά και επεκτάθηκε η λειτουργία στοιχειώδους μηχανής αναζήτησης σε σημασιολογικά δεδομένα β)Εμπλουτίστηκε ένα αρχικό σύνολο ταινιών με σημασιολογικές πληροφορίες οι οποίες χρησιμοποιήθηκαν για τη δημιουργία προφίλ χρηστών γ) Δημιουργήθηκαν κατάλληλα χαρακτηριστικά εκπαίδευσης του Ranking SVM έτσι ώστε να μάθει να ταξινομεί τα αποτελέσματα της αναζήτησης κάποιου χρήστη σύμφωνα με τις προτιμήσεις του δ) Πραγματοποιήθηκε πειραματική αξιολόγηση της μεθόδου.

**Λέξεις Κλειδιά:** << αναζήτηση με λέξεις κλειδιά, εξατομίκευση αναζήτησης, σημασιολογικά δεδομένα, Ranking SVM >>

Η σελίδα αυτή είναι σκόπιμα λευκή.

## Abstract

With the rise of the Semantic Web, the World Wide Web as we know it, a web of links between texts, is gradually becoming a web of data and meaning. This change has led to the development of search engines that extract information in rdf format. The growth of data available in the web, the different types of users and their needs as well as the ambiguities of the keyword search have created the need for personalized search.

In this thesis, we a) study the demands and restrictions in keyword search brought by the rdf model and extend the function of a simple semantic search engine b) infer user information through an initial set of film ratings, c) create new features for the Ranking SVM in order to train the system to rerank search results based on user interests d) perform experimental evaluation of our method.



**Keywords:** <<search personalization, rdf keyword search, Ranking SVM>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Γιώργο Γιαννόπουλο, διδακτορικό υπότροφο στο ΙΠΣΥ/Ε.Κ. "Αθηνά", για την καθοδήγηση και τις πολύτιμες υποδείξεις του στην εκπόνηση της παρούσας διπλωματικής εργασίας. Θα ήθελα επιπλέον να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Τιμολέοντα Σελλή που μου έδωσε την ευκαιρία να ασχοληθώ με ένα πολύ ενδιαφέρον θέμα.

Ευχαριστώ ιδιαίτερα την οικογένειά μου που με στήριξε σε όλη τη διάρκεια των σπουδών μου, καθώς και όλους όσους ήταν αυτή την περίοδο δίπλα μου και με το ενδιαφέρον τους, αλλά και την υπομονή τους, συνέβαλαν με τον τρόπο τους στην ολοκλήρωση της παρούσας εργασίας.

Η σελίδα αυτή είναι σκόπιμα λευκή.

# Πίνακας περιεχομένων

# 1

# *Εισαγωγή*

## *1.1 Η ανάγκη εξατομικευμένης αναζήτησης στο Σημασιολογικό Ιστό*

### *1.1.1 Εξατομίκευση Αναζήτησης*

Ο Παγκόσμιος Ιστός (World Wide Web) αποτελεί σήμερα μια τεράστια δεξαμενή πληροφοριών, το μέγεθος της οποίας αυξάνει καθημερινά. Χωρίς τις μηχανές αναζήτησης οι χρήστες θα έπρεπε για να ανακτήσουν οποιαδήποτε πληροφορία να πλοηγούνται στην πολύπλοκη δομή υπερσυνδέσμων (hyperlinks) του Ιστού, συχνά χάνοντας το στόχο τους. Οι μηχανές αναζήτησης διευκολύνουν το χρήστη παρέχοντάς του μια διεπαφή μέσω της οποίας μπορεί να πραγματοποιήσει ένα ερώτημα, να περιγράψει δηλαδή την πληροφορία που αναζητά. Στη συνέχεια, η μηχανή πραγματοποιεί μια αναζήτηση στα έγγραφα του Ιστού και επιστρέφει μια λίστα αποτελεσμάτων που πιθανώς περιέχουν τη ζητούμενη πληροφορία, ταξινομημένη από τα περισσότερο στα λιγότερο σχετικά αποτελέσματα.

Στα πρώτα χρόνια της χρήσης τους, οι μηχανές αναζήτησης επέστρεφαν αποτελέσματα με αποκλειστικό κριτήριο το ερώτημα που έθετε ο χρήστης. Έτσι, δύο διαφορετικοί χρήστες που έθεταν το ίδιο ερώτημα στην ίδια μηχανή αναζήτησης, θα λάμβαναν την ίδια λίστα αποτελεσμάτων. Ωστόσο, ο ρυθμός αύξησης των δεδομένων που εισάγονται καθημερινά

στον Ιστό σε συνδυασμό με τη δυνατότητα πρόσβασης σε αυτόν όλο και περισσότερων ανθρώπων με διαφορετικό πολιτιστικό υπόβαθρο, ανάγκες και ενδιαφέροντα, καθιστά ανεπαρκή αυτή την προσέγγιση και αναγκαία την επιπλέον παραμετροποίηση της αναζήτησης ώστε τα αποτελέσματά της να ανταποκρίνονται καλύτερα στις προτιμήσεις κάθε χρήστη. Είναι, δηλαδή, απαραίτητη η εξατομίκευση της αναζήτησης. Η εξατομίκευση στον Παγκόσμιο Ιστό είναι "η διαδικασία της συγκέντρωσης και αποθήκευσης πληροφοριών αναφορικά με τους χρήστες ενός website, η ανάλυση των πληροφοριών αυτών και, με βάση την ανάλυση, η αποστολή σε κάθε χρήστη της σωστής πληροφορίας στο σωστό χρόνο."

(Mulvenna et al., 2000)

Δεν είναι φυσικά μόνο οι μηχανές αναζήτησης που χρησιμοποιούν τεχνικές εξατομίκευσης. Σε κάθε πεδίο όπου ο πλούτος της διαθέσιμης πληροφορίας την καθιστά μη διαχειρίσιμη, είναι χρήσιμη ή και απαραίτητη η δυνατότητα εξόρυξης των δεδομένων που ενδιαφέρουν κάποιο συγκεκριμένο χρήστη εύκολα και γρήγορα. Έτσι, η εξατομίκευση βρίσκει εφαρμογή στο πεδίο των ηλεκτρονικών πωλήσεων, της ηλεκτρονικής μάθησης, των πληροφοριακών πυλών (με τη μορφή, παραδείγματος χάριν, παραμετροποίησης της αρχικής σελίδας του ιστότοπου), της προβολής πολυμεσικού περιεχομένου (ιστοσελίδες με μουσική, βίντεο) και γενικά οπουδήποτε το μέγεθος της συνολικής πληροφορίας μπορεί να λειτουργήσει αποθαρρυντικά για το χρήστη.

Ιδιαίτερα στην περίπτωση των μηχανών αναζήτησης, όμως, η εξατομίκευση δεν αποτελεί μόνο τρόπο να γίνει πιο ευχάριστη και ξεκούραστη η εμπειρία του χρήστη, αλλά τείνει να γίνει προϋπόθεση σωστής λειτουργίας. Ο λόγος για αυτό δεν είναι μόνο οι ανάγκες των διαφορετικών τύπων χρηστών ούτε η ραγδαία αύξηση των εγγράφων που καθημερινά εισάγονται στον Ιστό, αλλά και ο τρόπος αλληλεπίδρασης των χρηστών με τις μηχανές. Η πλέον διαδεδομένη μορφή ερωτημάτων είναι αυτά που σχηματίζονται από λίγες λέξεις-κλειδιά , είναι δηλαδή σύντομα και χωρίς δομή. Οι χρήστες επιλέγουν φράσεις που θεωρούν ότι περιγράφουν επαρκώς την πληροφορία που επιθυμούν να ανακτήσουν και οι μηχανές αναζήτησης αναζητούν την πληροφορία αυτή στην αχανή δομή εγγράφων του Παγκόσμιου Ιστού. Ο τρόπος αυτός σχηματισμού ερωτημάτων είναι που τα καθιστά εγγενώς διφορούμενα.

Ως ένα απλό παράδειγμα, το ερώτημα "Jaguar" αποτελείται από μια μοναδική λέξη κλειδί και είναι αντιπροσωπευτικό μιας αναζήτησης πληροφοριών για το συγκεκριμένο είδος πάνθηρα. Εύκολα αντιλαμβάνεται κανείς πως αντίστοιχα σχετικές μπορούν να θεωρηθούν πληροφορίες για γνωστή μάρκα αυτοκινήτων. Καθώς μάλιστα οι ιστοσελίδες που περιέχουν στοιχεία σχετικά με τα δύο αυτά είδη Jaguar είναι πάρα πολλές, το αποτέλεσμα που αναζητά ο χρήστης μπορεί να εμφανίζεται αρκετά χαμηλά στη λίστα αποτελεσμάτων. Ο συνδυασμός υψηλής ανάκλησης και μεγάλης ακρίβειας είναι μια από τις προκλήσεις στην τεχνολογία της

ανάκτησης πληροφορίας (information retrieval). Η αύξηση των λέξεων που χρησιμοποιούνται δεν είναι πάντα ικανή να φέρει πιο σχετικά αποτελέσματα και συχνά ο χρήστης καλείται να επαναδιατυπώσει το ερώτημά του με τρόπο σαφή και συγχρόνως όχι υπερβολικά περιοριστικό. Επειδή ο σκοπός ύπαρξης των μηχανών αναζήτησης είναι η διευκόλυνση του χρήστη, δεν μπορούμε να απαιτούμε από εκείνον να επαναδιατυπώνει συνεχώς τα ερωτήματά του μέχρι να βρει την ιστοσελίδα που καλύπτει τις ανάγκες του. Η εξατομίκευση της αναζήτησης στον Παγκόσμιο Ιστό είναι η κατεύθυνση των σύγχρονων μηχανών αναζήτησης .

### *1.1.2 Αναζήτηση στο Σημασιολογικό Ιστό*

Ο Σημασιολογικός Ιστός προωθείται από την Κοινοπραξία Παγκόσμιου Ιστού (World Wide Web Consortium, W3C), ένα διεθνή οργανισμό προτυποποίησης για τον Ιστό. Εμπνευστής της πρωτοβουλίας είναι ο Sir Tim Berners-Lee, ο άνθρωπος που επινόησε τον Παγκόσμιο Ιστό [LHL01].

Εμπλουτίζοντας τις ιστοσελίδες με σημασιολογικό περιεχόμενο, ο Σημασιολογικός Ιστός στοχεύει στο να μετατρέψει τον Παγκόσμιο Ιστό, έναν ιστό εγγράφων που αποτελείται από μη δομημένα ή ημιδομημένα δεδομένα, σε έναν Ιστό Δεδομένων, όπου η έννοια της πληροφορίας διαδραματίζει σημαντικότερο ρόλο. Σύμφωνα με το W3C, ο σημασιολογικός Ιστός θα προσφέρει μια κοινή πλατφόρμα για ανταλλαγή και επαναχρησιμοποίηση δεδομένων πέρα από τα όρια των επιμέρους εφαρμογών, επιχειρήσεων και κοινοτήτων.

Η εξέλιξη του Παγκόσμιου Ιστού στο Σημασιολογικό Ιστό μπορεί να αξιοποιηθεί από τις υπάρχουσες μηχανές αναζήτησης για την πιο ευφυή εκμετάλλευση της διαθέσιμης πληροφορίας. Οι μηχανές θα μπορούν πλέον να καταλάβουν τη σημασία της πληροφορίας που διαχειρίζονται . Για παράδειγμα, θα έχουν τη δυνατότητα να διαχωρίσουν το διαφορετικό σημασιολογικό περιεχόμενο στις φράσεις "Jennifer Lopez is single" και "CD single". Επιπλέον, οι μηχανές αναζήτησης θα μπορούν να λειτουργήσουν και ως μηχανές εξαγωγής συμπερασμάτων με χρήση επαγωγής στη διαθέσιμη γνώση.

Τα παραπάνω είναι δυνατά χάρη στην αναπαράσταση της γνώσης μέσω ενός γράφου διασυνδεδεμένων δεδομένων (Linked Data). Οι κόμβοι αντιπροσωπεύουν οντότητες (π.χ. πρόσωπα, επαγγέλματα, έννοιες, αντικείμενα, κατηγορίες) και οι ακμές περιγράφουν σχέσεις μεταξύ οντοτήτων. Το μοντέλο δεδομένων RDF (Resource Description Framework) περιγράφει το νέο τρόπο απόδοσης πληροφοριών σε μορφή τριπλέτας υποκείμενο-κατηγόρημα-αντικείμενο. Καθώς έχουν αναπτυχθεί γλώσσες περιγραφής σημασιολογίας και γλώσσες ερωτημάτων για να υποστηρίξουν το νέο πρότυπο, φαίνεται πως έχουν ωριμάσει οι συνθήκες για τη δημιουργία μηχανών αναζήτησης σε σημασιολογικά δεδομένα. Ωστόσο, οι

υπάρχουσες γλώσσες ερωτημάτων (π.χ. SPARQL) απαιτούν τη συμμόρφωση με συγκεκριμένους συντακτικούς κανόνες. Ζητούμενο για τις νέες μηχανές αναζήτησης, εκτός από την αποδοτική ανάκτηση πληροφοριών, είναι και η ευκολία στη χρήση ώστε στα οφέλη που προσφέρει ο νέος Ιστός να έχουν πρόσβαση και χρήστες έξω από την επιστημονική κοινότητα, οι οποίοι είναι περισσότερο εξοικειωμένοι με την αναζήτηση με λέξεις κλειδιά.

## *1.2 Αντικείμενο διπλωματικής*

Αντικείμενο της παρούσας διπλωματικής είναι η προσαρμογή και η αξιολόγηση ενός συστήματος εξατομίκευσης αναζήτησης σε σημασιολογικά δεδομένα. Η αναζήτηση πραγματοποιείται με λέξεις κλειδιά και ο χώρος αναζήτησης αποτελείται από τα αρχεία της DBpedia (http://dbpedia.org/About). Σκοπός της διπλωματικής είναι από ελάχιστες διαθέσιμες πληροφορίες σχετικά με τους χρήστες της μηχανής αναζήτησης να εξαχθούν δεδομένα τα οποία θα χρησιμοποιηθούν ως χαρακτηριστικά εκπαίδευσης του συστήματος εξατομίκευσης. Ζητούμενο είναι μετά την εκπαίδευση το σύστημα να μπορεί να ταξινομεί τα αποτελέσματα μιας αναζήτησης έτσι ώστε η σειρά τους να είναι περισσότερο αντιπροσωπευτική των ενδιαφερόντων του χρήστη.

### *1.2.1 Συνεισφορά*

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήσαμε δύο είδη συστημάτων εξατομίκευσης και επιλέξαμε το Ranking SVM ως το καταλληλότερο για τους σκοπούς της εργασίας μας.
2. Μελετήσαμε και επεκτείναμε την υπάρχουσα στοιχειώδη μηχανή αναζήτησης σε σημασιολογικά δεδομένα
3. Αντιστοιχίσαμε ταινίες από τα δεδομένα του Netflix Prize Award σε εγγραφές της DBpedia εμπλουτίζοντας τα αρχικά δεδομένα με σημασιολογικές πληροφορίες και δημιουργήσαμε έτσι προφίλ ενδιαφερόντων για χρήστες του Netflix
4. Δημιουργήσαμε χαρακτηριστικά για την εκπαίδευση του Ranking SVM προσαρμοσμένα στην εφαρμογή σε ταινίες, ηθοποιούς και σκηνοθέτες
5. Χρησιμοποιήσαμε το εκπαιδευμένο SVM στην εξατομικευμένη ταξινόμηση αποτελεσμάτων της μηχανής αναζήτησης ώστε η σειρά εμφάνισής τους να προσαρμόζεται στις προτιμήσεις κάθε χρήστη

## *1.3 Οργάνωση κειμένου*

Στο κεφάλαιο 2 εξετάζουμε τον τρόπο λειτουργίας δύο διαφορετικών μεθόδων εξατομίκευσης αναζήτησης σε rdf δεδομένα: την εξάπλωση βαρών στον rdf γράφο (propagation) και το Ranking SVM. Επίσης γίνεται αναφορά στα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν.

Στο κεφάλαιο 3 περιγράφουμε το αρχικό μας σύνολο δεδομένων, τη διαδικασία εμπλουτισμού του μέσω της DBpedia και τα κριτήρια επιλογής χρηστών για τα πειράματά μας.

Στο κεφάλαιο 4 εξηγούμε τους λόγους που απορρίψαμε τη χρήση του propagation στη χρήση εξατομίκευσης. Επίσης περιγράφεται ο τρόπος με τον οποίο εξήχθησαν πληροφορίες σχετικά με την ιεραρχία των δεδομένων του rdf γράφου που μας ενδιαφέρουν.

Στο κεφάλαιο 5 εξετάζουμε τα προβλήματα που έχουμε να αντιμετωπίσουμε στη χρήση του Ranking SVM εξαιτίας των περιορισμένων δεδομένων που διαθέτουμε. Εξηγούμε τις ανάγκες της προσέγγισής μας και περιγράφουμε αναλυτικά τα χαρακτηριστικά που δημιουργήσαμε για την εκπαίδευση του SVM.

Στο κεφάλαιο 6 περιγράφουμε τη λειτουργία της υπάρχουσας μηχανής αναζήτησης και τους τρόπους με τους οποίους την επεκτείναμε.

Στο κεφάλαιο 7 παρουσιάζονται και σχολιάζονται τα αποτελέσματα των πειραμάτων μας

Στο κεφάλαιο 8 κάνουμε μια ανασκόπηση των συμπερασμάτων που εξήχθησαν στα πλαίσια της διπλωματικής και κάνουμε αναφορά σε ανοιχτά ζητήματα, όπως μελλοντικές επεκτάσεις της μεθόδου μας.

# 2

# *Σχετικές εργασίες*

Στο κεφάλαιο αυτό παρουσιάζονται δύο διαφορετικές μέθοδοι εξατομικευμένης ταξινόμησης αποτελεσμάτων αναζήτησης που μελετήθηκαν στα πλαίσια της διπλωματικής. Ιδιαίτερη αναφορά γίνεται στους τρόπους με τους οποίους κάθε μέθοδος αξιοποιεί τις διαθέσιμες πληροφορίες σχετικά με τα ενδιαφέροντα και τις προτιμήσεις των χρηστών.

## *2.1 Μέθοδοι εξατομίκευσης αναζήτησης σε σημασιολογικά δεδομένα*

Η εξατομίκευση αναζήτησης σε σημασιολογικά δεδομένα απασχολεί τα τελευταία χρόνια την ερευνητική κοινότητα. Αναπόσπαστο κομμάτι της είναι η συλλογή πληροφοριών για το χρήστη και η αξιοποίησή τους για τη δημιουργία ενός προφίλ που να ενσωματώνει και να περιγράφει επιτυχώς τα ενδιαφέροντά του.

Γνωρίζουμε ότι οι χρήστες περιορίζονται στον έλεγχο των αποτελεσμάτων που βρίσκονται σχετικά ψηλά στη λίστα που τους επιστρέφει η μηχανή αναζήτησης. Είναι λοιπόν σημαντικό στις θέσεις αυτές να τοποθετηθούν τα αποτελέσματα που περιλαμβάνουν την πληροφορία που αναζητά ο χρήστης . Ζητούμενο δεν είναι μόνο τα αποτελέσματα να είναι σωστά με βάση το ερώτημα που τέθηκε, αλλά και κατάλληλα για το συγκεκριμένο χρήστη. Για να καταστεί αυτό δυνατό είναι απαραίτητη κάποιου είδους πληροφορία για τα πεδία που τον ενδιαφέρουν.

Από την άλλη πλευρά, οι χρήστες είναι συνήθως απρόθυμοι να παρέχουν ρητά πληροφορίες σχετικά με τις προτιμήσεις τους. Για το λόγο αυτό χρειαζόμαστε τεχνικές εξαγωγής συμπερασμάτων για τα ενδιαφέροντα του χρήστη από στοιχεία που δίνονται έμμεσα στο σύστημα.

Η αξιοποίηση του ιστορικού των αναζητήσεων είναι μια προφανής λύση. Ιστοσελίδες που ο χρήστης έχει επισκεφθεί στο παρελθόν ή, στην περίπτωση RDF δεδομένων, οντότητες (κόμβοι δηλαδή του RDF γράφου) που έχει προσπελάσει, προφανώς ανήκουν στα ενδιαφέροντά του και έχει νόημα να προωθούνται και σε μελλοντικές σχετικές ή και ακριβώς ίδιες αναζητήσεις. Δεν έχει ιδιαίτερη αξία βέβαια η δυνατότητα εξατομίκευσης μόνο σε αναζητήσεις που έχουν επαναληφθεί στο παρελθόν ή σε αποτελέσματα που γνωρίζουμε με βεβαιότητα ότι ενδιαφέρουν το χρήστη. Σκοπός μας είναι, αξιοποιώντας το ιστορικό του, να προβλέπουμε πόσο ανταποκρίνονται στις ανάγκες του πληροφορίες που δεν έχει ξανασυναντήσει καθώς και να ερμηνεύουμε τα ερωτήματά του όταν αυτά δεν είναι διατυπωμένα με αρκετή σαφήνεια. Οι εφαρμογές που αναπτύσσονται με αυτό το στόχο διαχωρίζονται με κριτήριο τον τρόπο που αποτυπώνουν τη διαθέσιμη για το χρήστη πληροφορία και "εκπαιδεύονται " να αξιολογούν τα αποτελέσματα με βάση τις προτιμήσεις του. Θα αναφέρουμε εδώ δύο κατηγορίες τέτοιων εφαρμογών. Στην πρώτη περιλαμβάνονται εφαρμογές που μοντελοποιούν τα ενδιαφέροντα του χρήστη ως υπογράφους του συνολικού γράφου δεδομένων και χρησιμοποιούν την εξάπλωση βαρών σε αυτόν για να βρουν ποια τμήματά του αφορούν το χρήστη. Στη δεύτερη κατηγορία βρίσκονται εφαρμογές που κάνουν χρήση μηχανικής μάθησης για να εκπαιδεύσουν ένα νευρωνικό δίκτυο να προβλέπει τις προτιμήσεις του χρήστη.

Στο σημείο αυτό δε μας ενδιαφέρει ο τρόπος με τον οποίο αποκτήθηκαν οι πληροφορίες για το χρήστη ούτε το είδος των αποτελεσμάτων που παρουσιάζονται από τη μηχανή αναζήτησης, παρά μόνο ο τρόπος με τον οποίο αυτά αξιολογούνται σε σχέση με το ιστορικό του χρήστη.

### *2.1.1 Εξάπλωση βαρών σε γράφους (Propagation)*

Η αναπαράσταση μέσω γραφημάτων αποτελεί μια φυσιολογική επιλογή για τα σημασιολογικά δεδομένα. Ο πρώτος όρος (υποκείμενο-subject) και ο τρίτος όρος (αντικείμενο-object) κάθε rdf τριπλέτας αναπαρίσταται με έναν κόμβο στο γράφο, ενώ ο ενδιάμεσος όρος (ιδιότητα - predicate) από μια ακμή που ενώνει το υποκείμενο με το αντικείμενο. Οι εφαρμογές που χρησιμοποιούν την τεχνική της διάδοσης βαρών μέσα σε ένα γράφο, εκμεταλλεύονται αυτόν ακριβώς τον τρόπο αναπαράστασης των δεδομένων.

Η γενική ιδέα είναι πως η συνολική γνώση (το dataset) σχηματίζει ένα γράφο. Σαν είσοδο στο σύστημα εξατομίκευσης έχουμε κάποιους κόμβους ή και ακμές του γράφου στα οποία έχει αντιστοιχηθεί ένας βαθμός (βάρος) με βάση το ιστορικό του χρήστη. Στη συνέχεια ένας αλγόριθμος διάδοσης βαρών ξεκινά από καθέναν από τους κόμβους (και τις ακμές) της εισόδου και διασχίζει το γράφο δίνοντας κατάλληλα υπολογισμένα βάρη σε νέες ακμές και κόμβους. Όταν ολοκληρωθεί η διαδικασία, τα κομμάτια του αρχικού γράφου που έχουν αποκτήσει βάρη σκιαγραφούν τα πεδία ενδιαφέροντος του χρήστη. Ανάλογα μάλιστα με το πόσο μεγάλα ή μικρά είναι τα βάρη σε κάποιους υπογράφους, διαχωρίζουμε τα ενδιαφέροντα του χρήστη σε περισσότερο ή λιγότερο σημαντικά.

Δίνουμε τώρα συνοπτικά κάποια σημεία στα οποία διαφοροποιούνται οι εφαρμογές αυτής της τεχνικής που συναντάμε στη βιβλιογραφία καθώς και τα σημεία που πρέπει να προσεχθούν ώστε ο αλγόριθμος να συγκλίνει , δηλαδή τα βάρη που έχουν αποδοθεί στα διάφορα τμήματα του αρχικού γράφου να είναι αντιπροσωπευτικά του χρήστη [Dud08], [SMB07], [CLO11], [RSP04], [JT09], [DLB09], [CP11], [DEL+08], [Sieg+07].

- Απόδοση βαρών μόνο σε κόμβους ή και σε ακμές

  Η επιλογή εξαρτάται από το είδος της διαθέσιμης πληροφορίας για το χρήστη και εάν σε αυτή, εκτός από προτιμήσεις για οντότητες και κλάσεις, συμπεριλαμβάνονται προτιμήσεις για το είδος των σχέσεων που συνδέουν μεταξύ τους αυτές τις οντότητες, οι οποίες μπορούν πιθανώς να εξαχθούν από τη συχνότητα με την οποία ένας χρήστης επιλέγει να ανακτήσει τέτοιου είδους σημασιολογικές σχέσεις.

  Στην περίπτωση που αυτή η πληροφορία είναι διαθέσιμη, παρουσιάζει επιπλέον ενδιαφέρον το πως μπορούμε να ανιχνεύσουμε σημασιολογική ομοιότητα μεταξύ δύο διαφορετικών σχέσεων (ακμών), ώστε εκτός από τις σχέσεις για τις οποίες ο χρήστης έχει άμεσα δείξει ενδιαφέρον, να εξαπλωθεί το ενδιαφέρον του και σε παρόμοιου τύπου σχέσεις (π.χ. οι σχέσεις "born in" και "originates from").

- Επιλογή ακμών μέσω των οποίων θα επιτρέπεται η εξάπλωση βαρών

  Δεδομένου ότι μια τυπική βάση rdf δεδομένων μπορεί να περιλαμβάνει μεγάλο αριθμό διαφορετικών σχέσεων οι οποίες αντιπροσωπεύουν διαφορετικού είδους ακμές στο γράφο γνώσης, έχει σημασία να επιλεγούν οι ακμές εκείνες μέσω των οποίων η διάδοση βαρών έχει νόημα και δεν προκαλεί διασπορά του αρχικού ενδιαφέροντος του χρήστη σε κόμβους που σημασιολογικά απέχουν σημαντικά από τον αρχικό κόμβο. Διακρίνουμε δύο βασικές μεθόδους:

  - εξάπλωση μόνο μέσω σχέσεων κλάσης-υποκλάσης (κάθετη εξάπλωση)

- εξάπλωση μέσω σχέσεων κλάσης-υποκλάσης, αλλά και άλλων επιλεγμένων σχέσεων

- Ποσοστό βάρους που εξαπλώνεται σε κάθε βήμα

  Η αρχική μας πληροφορία αποτελείται από το ενδιαφέρον του χρήστη για κάποιο κόμβο του rdf γράφου, το οποίο μεταφράζεται σε έναν αριθμό και αντιπροσωπεύεται από το βάρος του κόμβου για τον οποίο ο χρήστης άμεσα έχει εκφράσει αυτή την προτίμηση. Η προτίμηση αυτή μεταφράζεται σε έμμεση προτίμηση για τους γειτονικούς κόμβους του αρχικού, μέσω της εξάπλωσης του βάρους του σε αυτούς. Είναι σημαντικό, λοιπόν, να εξετάσουμε αν αυτό το βάρος εξαπλώνεται ίδιο ή υφίσταται κάποια μείωση καθώς απομακρυνόμαστε από τον αρχικό κόμβο. Η επιλογή του κατάλληλου ποσοστού που θα εξαπλωθεί επηρεάζεται από τους εξής παράγοντες:

  - μήκος μονοπατιού (απόσταση) από τον αρχικό κόμβο
  - είδος ακμής μέσω της οποίας γίνεται η εξάπλωση
  - θέση της κλάσης του κόμβου στην ιεραρχία κλάσεων
  - επιβράβευση κόμβων που δέχονται μέσω εξάπλωσης βάρος από πολλούς απογόνους τους
  - αποφυγή δυσανάλογης εξάπλωσης βάρους από έναν κόμβο που ο χρήστης επισκέπτεται/προσπελαύνει συχνά, αλλά δεν έχει δείξει άμεσα ενδιαφέρον για τους συγγενικούς/γειτονικούς του κόμβους

- Κριτήρια τερματισμού εξάπλωσης

  Προς αποφυγή εξάπλωσης σε μη σχετικά κομμάτια του rdf γράφου, είναι σημαντικό να οριστούν τα κριτήρια με τα οποία τερματίζεται η διαδικασία της εξάπλωσης του βάρους από τον αρχικό κόμβο. Μερικά από αυτά είναι:

  - κόμβοι συγκεκριμένου τύπου (π.χ. κλάση) στους οποίους σταματάμε
  - μήκος μονοπατιού (απόσταση) από τον αρχικό κόμβο
  - κατώφλι βάρους
  - πλήθος ακμών που ξεκινούν από τον κόμβο στον οποίο βρισκόμαστε (fanout)

- Επίδραση ενός νέου βάρους και της εξάπλωσής του στα υπάρχοντα βάρη

  Η μείωση των προηγούμενων βαρών καθώς εισάγονται στο σύστημα νέα βάρη που αντιπροσωπεύουν νέα ενδιαφέροντα του χρήστη αφενός επιβάλλει ένα όριο στην

αύξηση των βαρών βοηθώντας στη σύγκλιση της μεθόδου και αφετέρου ενσωματώνει τη χρονική εξέλιξη των ενδιαφερόντων του χρήστη. Επιπλέον, η προσαρμογή των βαρών και της εξάπλωσής τους ανάλογα με το πλήθος των επαναλήψεων της διαδικασίας εξασφαλίζει ότι η προσπέλαση ενός νέου κόμβου από το χρήστη επηρεάζει όλο και λιγότερο τα βάρη που έχουν ανατεθεί, όσο μεγαλώνει ο όγκος της πληροφορίας που έχει αξιοποιηθεί σε προηγούμενες επαναλήψεις (δηλαδή ένα νέο βάρος θα επηρεάσει λιγότερο τη διαδικασία αν διαθέτουμε ήδη αρκετές πληροφορίες για τα ενδιαφέροντα του χρήστη).

Οι παραπάνω επιλογές καθορίζουν το είδος των αποτελεσμάτων που η μηχανή αναζήτησης μπορεί να αξιολογήσει και πρέπει να εξασφαλίζουν ότι το προφίλ του χρήστη συγκλίνει τελικά σε κάποια τμήματα του γράφου γνώσης, ότι υπάρχουν δηλαδή κόμβοι και ακμές που σταθερά συγκεντρώνουν υψηλές βαθμολογίες (βάρη) οι οποίες πράγματι αντιπροσωπεύουν τα ενδιαφέροντα του χρήστη και το σύστημα δεν επηρεάζεται υπερβολικά από μια νέα βαθμολογία.

Στη συνέχεια, αφού η μηχανή αναζήτησης ανακτήσει τα αποτελέσματα για ένα ερώτημα, τα αξιολογεί και τα αναδιατάσσει χρησιμοποιώντας τα βάρη που έχουν υπολογιστεί στο προηγούμενο στάδιο.

### *2.1.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines- SVM)*

Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την αυτόματη κατασκευή ενός μοντέλου ταξινόμησης. Σαν είσοδο στο σύστημα μηχανικής μάθησης δίνουμε ερωτήματα και ταξινομημένες λίστες αποτελεσμάτων για καθένα από αυτά. Το σύστημα χρησιμοποιεί αυτά τα δεδομένα για να μάθει μια συνάρτηση ταξινόμησης η οποία στη συνέχεια θα εφαρμόζεται στις λίστες καινούργιων αποτελεσμάτων. Από τα διάφορα συστήματα μηχανικής μάθησης, οι μηχανές διανυσμάτων υποστήριξης (SVM) αποδεικνύονται ιδιαίτερα αποτελεσματικές σε τέτοιες εφαρμογές.

Ένα σύνολο ερωτημάτων με τα ταξινομημένα με βάση τις προτιμήσεις του χρήστη αποτελέσματα για το καθένα αποτελούν τα δεδομένα εκπαίδευσης του συστήματος. Ζητούμενο είναι να βρούμε κατάλληλα χαρακτηριστικά των αποτελεσμάτων που να αποτυπώνουν τις μεταξύ τους διαφορές τόσο γενικά όσο και στο πλαίσιο του συγκεκριμένου ερωτήματος. Θα δώσουμε τώρα μερικές κατηγορίες χαρακτηριστικών που συναντάμε στη βιβλιογραφία, μαζί με συγκεκριμένα παραδείγματα για καθεμία [MBH+09], [DF11], [DFTM12], [LLNF12].

- Χαρακτηριστικά που προκύπτουν από το ιστορικό του χρήστη
  - πλήθος εμφανίσεων του ονόματος (label) ενός κόμβου Κ σε ερώτημα του χρήστη (ή επιπλέον τι ποσοστό λέξεων του ερωτήματος σχετίζονται με το όνομα του Κ)
  - πλήθος φορών που ένας κόμβος Κ ανακτήθηκε ως αποτέλεσμα κάποιου ερωτήματος
- Χαρακτηριστικά που εκφράζουν την ομοιότητα ερωτήματος-αποτελέσματος
  - πλήθος κοινών όρων στο ερώτημα και στο όνομα (label) του κόμβου Κ αν το αποτέλεσμα είναι ένας κόμβος

- Χαρακτηριστικά που προκύπτουν από τη δομή του rdf γράφου και αφορούν ένα αποτέλεσμα-κόμβο Κ
  - πλήθος εισερχόμενων ακμών (σχέσεις στις οποίες ο κόμβος είναι αντικείμενο)

    Επεκτείνεται για μονοπάτια μεταβλητού μήκους που καταλήγουν στον Κ.
  - πλήθος εξερχόμενων ακμών (σχέσεις στις οποίες ο κόμβος είναι υποκείμενο)

    Επεκτείνεται για μονοπάτια μεταβλητού μήκους που ξεκινούν από τον Κ.
  - πλήθος σχέσεων στις οποίες ο κόμβος είναι υποκείμενο και το αντικείμενο είναι literal
  - πλήθος διαφορετικού τύπου ακμών που εξέρχονται από τον Κ

    Επεκτείνεται για ακμές ν βήματα μετά τον Κ.
  - πλήθος διαφορετικού τύπου ακμών που εισέρχονται στον Κ

    Επεκτείνεται για ακμές ν βήματα πριν τον Κ.
  - μέση συχνότητα εμφάνισης στον rdf γράφο των ακμών που εισέρχονται στον Κ

    Επεκτείνεται για ακμές ν βήματα πριν τον Κ.
  - μέση συχνότητα εμφάνισης στον rdf γράφο των ακμών που εξέρχονται από τον Κ
  - πλήθος κατηγοριών (π.χ. Wikipedia κατηγορίες) στις οποίες ανήκει
  - μέγεθος της μεγαλύτερης κατηγορίας στην οποία ανήκει
  - μέγεθος της μικρότερης κατηγορίας στην οποία ανήκει
  - μέγεθος της μεσαίας(από πλευράς μεγέθους) κατηγορίας στην οποία ανήκει
  - βάθος του κόμβου στο γράφο των δεδομένων (εξαρτάται από την οντολογία που έχει χρησιμοποιηθεί και εκφράζει τη θέση της κλάσης του κόμβου στην ιεραρχία κλάσεων)

- βάθος στην ιεραρχία κατηγοριών SKOS
- πλήθος σελίδων που επανακατευθύνουν (redirection pages) στον K

- Χαρακτηριστικά που προκύπτουν από τη δομή του rdf γράφου και αφορούν ένα αποτέλεσμα-μονοπάτι
  - πλήθος κόμβων που περιλαμβάνει
  - πλήθος πολύπλοκων ιδιοτήτων που εμφανίζονται (ιδιοτήτων δηλαδή που έχουν δικές τους ιδιότητες)
  - εφαρμογή χαρακτηριστικών που αναφέρθηκαν παραπάνω για ξεχωριστούς κόμβους και χρήση μέσου όρου των τιμών που προκύπτουν για το σύνολο των κόμβων που περιλαμβάνει

### *2.1.3 Επιλογή μεθόδου εξατομίκευσης*

Για λόγους που θα αναλυθούν σε επόμενα κεφάλαια, από τις δύο προαναφερθείσες μεθόδους εξατομίκευσης αναζήτησης, στην παρούσα εργασία θα χρησιμοποιηθεί μόνο η δεύτερη, δηλαδή το Ranking SVM.

## *2.2 Εργαλεία*

### *2.2.1 Apache Lucene*

Το Apache Lucene (http://lucene.apache.org/) είναι μια ελεύθερου/ανοιχτού κώδικα βιβλιοθήκη που παρέχει λογισμικό για ανάκτηση πληροφοριών και χρησιμοποιείται ευρέως σε εφαρμογές που απαιτούν σύνταξη ευρετηρίου (indexing) και αναζήτηση σε κείμενα. Ειδικότερα, είναι πολύτιμη στην υλοποίηση μηχανών αναζήτησης στο Internet, αλλά και τοπικά για αναζήτηση σε έναν ιστότοπο. Κεντρική ιδέα για τη λειτουργία της είναι η δημιουργία εγγράφων (documents) που περιέχουν πεδία (fields) κειμένου.

Στην εφαρμογή μας, χρησιμοποιούνται μέθοδοι του Lucene αρχικά για τη σύνταξη ευρετηρίων των δεδομένων του χώρου αναζήτησης και στη συνέχεια για τις αναζητήσεις που πραγματοποιεί η μηχανή αναζήτησης μετά την υποβολή ενός ερωτήματος από κάποιον χρήστη.

### *2.2.2 Apache Jena*

Το Apache Jena (http://jena.apache.org/) είναι ένα Java framework που παρέχει μια συλλογή εργαλείων και βιβλιοθηκών σε Java για την ανάπτυξη εφαρμογών Σημασιολογικού Ιστού και διασυνδεδεμένων δεδομένων. Σε αυτό περιλαμβάνονται μέθοδοι για την αποδοτική αποθήκευση μεγάλου αριθμού από RDF δεδομένα (τριπλέτες) στο δίσκο, καθώς και για την εκτέλεση SPARQL ερωτημάτων σε αυτά. Τα παραπάνω χαρακτηριστικά αξιοποιούνται από τη μηχανή αναζήτησης για την εύρεση σχέσεων που να συνδέουν μεταξύ τους κόμβους του rdf γράφου.

### *2.2.3 SVM$^{rank}$*

Το SVM$^{rank}$ είναι μια υλοποίηση μηχανής διανυσμάτων υποστήριξης (Support Vector Machine) γραμμένη σε γλώσσα C από τον Thorsten Joachims (http://www.cs.cornell .edu/People/tj/). Η συγκεκριμένη υλοποίηση είναι κατάλληλη για εφαρμογές εξατομίκευσης αποτελεσμάτων αναζήτησης. Σαν είσοδος για την εκπαίδευση της μηχανής δίνεται μια σειρά ερωτημάτων καθώς και μια ταξινομημένη λίστα αποτελεσμάτων για καθένα από αυτά. Τα αποτελέσματα δίνονται με τη μορφή ν-διάστατων διανυσμάτων, όπου ν ο αριθμός των χαρακτηριστικών που έχουν επιλεγεί ως κατάλληλα για την περιγραφή των αποτελεσμάτων. Ο τρόπος λειτουργίας του αλγόριθμου εκπαίδευσης από τέτοιου τύπου δεδομένα περιγράφεται στο [Joa06].

# 3

# *Προεπεξεργασία Δεδομένων*

Για την εφαρμογή μεθόδων εξατομίκευσης αναζήτησης, χρειαζόμαστε ένα σύνολο χρηστών και πληροφορίες σχετικά με τα ενδιαφέροντα και τις προτιμήσεις τους. Για το σκοπό αυτό, χρησιμοποιούμε τα αρχεία από το διαγωνισμό Netflix Prize (http://www.netflixprize.com/), τα οποία περιέχουν πληροφορίες για βαθμολογίες 17770 ταινιών και σειρών από χρήστες καθώς και τις ημερομηνίες που οι χρήστες έδωσαν αυτές τις βαθμολογίες. Τα αρχεία προορίζονταν να βοηθήσουν στην βελτίωση του αλγόριθμου συνεργατικού φιλτραρίσματος (collaborative filtering) που χρησιμοποιεί το Netflix για να προτείνει στους χρήστες του ταινίες που πιθανώς να τους αρέσουν με βάση τις προτιμήσεις χρηστών με παρόμοια ενδιαφέροντα. Ο στόχος αυτός είναι αρκετά διαφορετικός από το δικό μας και η διαθέσιμη πληροφορία σαφώς περιορισμένη σε σχέση με τα στοιχεία που χρειαζόμαστε για να δημιουργήσουμε ένα αντιπροσωπευτικό προφίλ για τους χρήστες , ικανό να περιγράψει επιλογές που δεν αφορούν αποκλειστικά ταινίες. Ωστόσο, δεδομένου ότι συχνά οι χρήστες είναι απρόθυμοι να παρέχουν άμεσα πληροφορίες σχετικά με τις προτιμήσεις τους, σκοπός μας είναι να δούμε κατά πόσο μπορούμε να επεκταθούμε από τις βαθμολογίες ταινιών σε προτιμήσεις τους σε πεδία όπως ηθοποιοί, βραβεία , σκηνοθέτες, θεματικές ενότητες και είδη ταινιών. Σε κάθε περίπτωση, βέβαια, τα ερωτήματα στη μηχανή αναζήτησης θα πρέπει να κινούνται στο ευρύτερο πεδίο του κινηματογράφου, αφού δεν έχουμε κατάλληλα δεδομένα για εξατομίκευση αποτελεσμάτων μιας γενικού περιεχομένου αναζήτησης .

Στο παρόν κεφάλαιο παρουσιάζεται αναλυτικά το είδος των πληροφοριών που διαθέτουμε καθώς και ο τρόπος με τον οποίο συνδυάστηκαν ώστε να σχηματισθεί ένα σύνολο δεδομένων κατάλληλο για τους σκοπούς της εργασίας μας.

## *3.1 Εμπλουτισμός δεδομένων Netflix με περισσότερες πληροφορίες*

Για να μπορέσουν να χρησιμοποιηθούν προς την επιθυμητή κατεύθυνση τα δεδομένα από το Netflix Prize, χρειάζεται να εμπλουτιστούν με περισσότερες πληροφορίες, αφού το Netflix παρέχει μόνο τίτλο και έτος κυκλοφορίας για κάθε ταινία. Θα χρησιμοποιήσουμε για το λόγο αυτό τα αρχεία της DBpedia, μιας από τις πλουσιότερες βάσεις δομημένων δεδομένων που αριθμεί σχεδόν 1 δισεκατομμύριο τριπλέτες . Το περιεχόμενο της DBpedia εξάγεται από τη Wikipedia, αναπαρίσταται σε RDF μορφή και κατηγοριοποιείται με τη χρήση μιας OWL οντολογίας. Γνωρίζουμε ότι στα αρχεία της περιλαμβάνονται εγγραφές για περίπου 60000 ταινίες.

Αρχικά πρέπει να αντιστοιχίσουμε τις ταινίες από τα αρχεία του Netflix σε εγγραφές της DBpedia. Χρησιμοποιούμε την έκδοση 3.7 των αρχείων της DBpedia (http://wiki. dbpedia.org/Downloads37). Για να κάνουμε την αντιστοίχιση των τίτλων χρησιμοποιούμε αυστηρό ταίριασμα συμβολοσειρών, μετά από την αφαίρεση-αντικατάσταση ειδικών χαρακτήρων, σημείων στίξης και κενών. Απαιτούμε εκτός από τους τίτλους να ταυτίζεται και το έτος κυκλοφορίας της ταινίας. Την πληροφορία για το έτος κυκλοφορίας μιας ταινίας στη DBpedia εξάγουμε είτε από την ιδιότητα <http://dbpedia.org/ontology/releaseDate>, όπου αυτή είναι διαθέσιμη, είτε από το label της ταινίας όταν το έτος περιλαμβάνεται σε αυτό (π.χ. "The Raven (1963 film)" για την ταινία με URI http://dbpedia.org/page/The_ Raven_%281963_film%29) . Γνωρίζουμε ότι και στο Netflix και στη DBpedia η πληροφορία αυτή δεν είναι πάντα ακριβής με αποτέλεσμα πιθανώς τη λανθασμένη απόρριψη κάποιων αντιστοιχίσεων. Ο μεγάλος αριθμός των ταινιών με ίδιο τίτλο, όμως, κάνει απαραίτητη τη χρήση και του έτους για τη διασφάλιση της ορθότητας των αντιστοιχίσεων. Δεδομένου ότι, όπως είπαμε, το προφίλ ενός χρήστη θα σχηματισθεί μέσω ελλιπούς πληροφορίας, σκοπός μας είναι να ελαχιστοποιήσουμε τα σφάλματα που μπορεί να δημιουργήσουν επιπλέον προβλήματα στη συνοχή του προφίλ, ώστε να παρέχουμε στο σύστημα εξατομίκευσης όσο γίνεται πιο πλήρη και ακριβή δεδομένα εισόδου.

Με τον τρόπο που περιγράφηκε καταφέρνουμε να αντιστοιχίσουμε 5179 ταινίες. Ο αριθμός αυτός είναι επαρκής για τις ανάγκες των πειραμάτων μας, οπότε δε χρειάστηκε να

καταφύγουμε σε μεθόδους μερικού ταιριάσματος συμβολοσειρών (approximate string matching). Μετά την αντιστοίχιση, εκτός από το έτος κυκλοφορίας διαθέτουμε πλήθος στοιχείων για κάθε ταινία, όπως το σκηνοθέτη, τους ηθοποιούς, το σεναριογράφο και τη χώρα προέλευσης . Από τα δεδομένα της DBpedia απουσιάζει ωστόσο το είδος στο οποίο ανήκει μια ταινία. Είναι έμμεσα διαθέσιμο από τις SKOS κατηγορίες και την οντολογία YAGO, για τα οποία θα μιλήσουμε σε επόμενο κεφάλαιο, αλλά σε πιο πολύπλοκη μορφή. Για την απόκτηση αυτής της πληροφορίας χρησιμοποιήσαμε τα δεδομένα του IMDB, διαθέσιμα στο σύνδεσμο ftp://ftp.fu-berlin.de/pub/misc/movies/database/. Και πάλι για να αντιστοιχίσουμε τις ταινίες από το ένα σύνολο σε ταινίες του άλλου, χρησιμοποιήσαμε την ίδια τεχνική με παραπάνω και ανακτήσαμε πληροφορίες για το είδος 1752 ταινιών. Κάθε ταινία από το IMDB μπορεί να ανήκει σε ένα ή περισσότερα από τα συνολικά 29 είδη. Για κάθε είδος στο οποίο κατατάσσεται μια ταινία, δημιουργούμε μια σχέση της μορφής <URI ταινίας> <Genre> <Είδος στο οποίο ανήκει η ταινία> και την εισάγουμε στο σύνολο των δεδομένων μας.

Σημειώνουμε πως ο χαρακτηρισμός μιας ταινίας από πολλά είδη είναι συνήθης πρακτική σε σχετικές εφαρμογές και περισσότερο θεωρείται ότι εμπλουτίζει τις πληροφορίες παρά δημιουργεί ασάφειες.

Στο σημείο αυτό πρέπει να αναφέρουμε ότι για τη δημιουργία μιας όσο το δυνατόν πληρέστερης δεξαμενής πληροφοριών σχετικά με τις ταινίες που μας ενδιαφέρουν, εξετάσαμε το ενδεχόμενο να συμπεριλάβουμε τα στοιχεία από το LinkedMDB (http:// data.linkedmdb.org/), τη διασυνδεδεμένη μορφή του IMDB. Η μόνη επιπλέον διαθέσιμη πληροφορία που παρέχεται από αυτή τη βάση δεδομένων συγκριτικά με τη DBpedia αφορά συγκεκριμένους ρόλους που κατείχαν οι ηθοποιοί στις ταινίες που συμμετείχαν, όπως το όνομα κάποιου χαρακτήρα της ταινίας. Καθώς δεν κρίθηκε σκόπιμο να συμπεριληφθεί αυτή η πληροφορία στο σύνολο δεδομένων όπου θα πραγματοποιείται η αναζήτηση, όλες οι άλλες πληροφορίες μας εκτός από το είδος, προέρχονται από τη DBpedia.

## *3.2 Στατιστικά χαρακτηριστικά συνόλου δεδομένων*

Έχοντας καταλήξει σε 5179 ταινίες, πραγματοποιούμε κάποιους στατιστικούς ελέγχους ώστε να διαπιστώσουμε τι είδους πληροφορίες διαθέτουμε και πόσο πλήρεις είναι. Τα εικονιζόμενα στοιχεία των πινάκων 1-6 προέρχονται από το περιεχόμενο του αρχείου Ontology Infobox Properties.

Στον πρώτο πίνακα βλέπουμε στατιστικά στοιχεία για το σύνολο των σχέσεων στις οποίες οι ταινίες που μας ενδιαφέρουν έχουν θέση υποκείμενου. Η πρώτη στήλη περιέχει το όνομα της

ιδιότητας (κατηγόρημα-predicate) της σχέσης, η δεύτερη στήλη περιέχει τον αριθμό των εμφανίσεων αυτής της ιδιότητας στο σύνολο των δεδομένων και η τρίτη στήλη τον αριθμό των ταινιών για τις οποίες εμφανίζεται αυτή η ιδιότητα. Οι τιμές των δύο στηλών διαφέρουν, καθώς μια ιδιότητα μπορεί να εμφανίζεται για την ίδια ταινία περισσότερες από μια φορά (ίδιο υποκείμενο, ίδιο κατηγόρημα, διαφορετικό αντικείμενο). Η τέταρτη στήλη εκφράζει το ποσοστό των ταινιών που διαθέτουν την εν λόγω ιδιότητα, ενώ η πέμπτη στήλη το μέσο αριθμό εμφανίσεων της ιδιότητας στο σύνολο των 5179 ταινιών.

| **Ιδιότητα (dbp = http://dbpedia.org/ontology/)** | **Πλήθος εμφανίσεων ιδιότητας** | **Πλήθος ταινιών στις οποίες εμφανίζεται** | **Ποσοστό ταινιών στις οποίες εμφανίζεται** | **Μέσος αριθμός εμφανίσεων ανά ταινία** |
|---|---|---|---|---|
| dbp:starring | 23861 | 5038 | 97.28 | 4.61 |
| dbp:writer | 7984 | 4529 | 87.45 | 1.54 |
| dbp:producer | 6867 | 4078 | 78.74 | 1.33 |
| dbp:releaseDate | 5513 | 4770 | 92.10 | 1.06 |
| dbp:director | 5212 | 4937 | 95.33 | 1.01 |
| http://xmlns.com/foaf/0.1/name | 5212 | 5147 | 99.38 | 1.01 |
| dbp:distributor | 5160 | 4407 | 85.09 | 1.00 |
| dbp:runtime | 5134 | 4990 | 96.35 | 0.99 |
| dbp:musicComposer | 4837 | 3976 | 76.77 | 0.93 |
| dbp:cinematography | 3066 | 2916 | 56.30 | 0.59 |
| dbp:editing | 2793 | 2421 | 46.75 | 0.54 |
| dbp:language | 2481 | 2097 | 40.49 | 0.48 |
| dbp:gross | 2078 | 2008 | 38.77 | 0.40 |
| dbp:budget | 1856 | 1846 | 35.64 | 0.36 |
| dbp:country | 1812 | 1600 | 30.89 | 0.35 |
| dbp:basedOn | 195 | 136 | 2.63 | 0.04 |
| dbp:narrator | 190 | 182 | 3.51 | 0.04 |

**Πίνακας 1: Σχέσεις όπου οι ταινίες είναι υποκείμενα**

Ακολουθούν αντίστοιχα στοιχεία για τις ιδιότητες των σχέσεων στις οποίες οι ταινίες είναι αντικείμενα. Στον πίνακα περιλαμβάνονται οι ιδιότητες που εμφανίζονται τουλάχιστον δύο φορές στο σύνολο των σχέσεων.

| **Ιδιότητα (dbp = http://dbpedia.org/ontology/)** | **Πλήθος εμφανίσεων ιδιότητας** | **Πλήθος ταινιών στις οποίες εμφανίζεται** | **Ποσοστό ταινιών στις οποίες εμφανίζεται** | **Μέσος αριθμός εμφανίσεων ανά ταινία** |
|---|---|---|---|---|
| dbp:firstAppearance | 106 | 48 | 11.54 | 0.25 |
| dbp:award | 99 | 55 | 13.22 | 0.24 |
| dbp:previousWork | 87 | 84 | 20.19 | 0.21 |
| dbp:subsequentWork | 71 | 64 | 15.38 | 0.17 |
| dbp:knownFor | 34 | 34 | 8.17 | 0.08 |
| dbp:notableWork | 30 | 29 | 6.97 | 0.07 |
| dbp:basedOn | 30 | 30 | 7.21 | 0.07 |
| dbp:album | 29 | 28 | 6.73 | 0.07 |
| dbp:series | 24 | 16 | 3.85 | 0.06 |
| dbp:lastAppearance | 23 | 18 | 4.33 | 0.06 |
| dbp:product | 22 | 20 | 4.81 | 0.05 |
| dbp:related | 15 | 15 | 3.61 | 0.04 |
| dbp:associatedBand | 10 | 9 | 2.16 | 0.02 |
| dbp:associatedMusicalArtist | 10 | 9 | 2.16 | 0.02 |
| dbp:portrayer | 9 | 8 | 1.92 | 0.02 |
| dbp:musicComposer | 6 | 3 | 0.72 | 0.01 |
| dbp:producer | 5 | 4 | 0.96 | 0.01 |
| dbp:director | 5 | 5 | 1.20 | 0.01 |

| | | | | |
|---|---|---|---|---|
| dbp:starring | 4 | 4 | 0.96 | 0.01 |
| dbp:writer | 2 | 2 | 0.48 | 0.00 |
| dbp:billed | 2 | 2 | 0.48 | 0.00 |
| dbp:battle | 2 | 2 | 0.48 | 0.00 |
| dbp:spouse | 2 | 2 | 0.48 | 0.00 |
| dbp:openingTheme | 2 | 2 | 0.48 | 0.00 |
| dbp:distributor | 2 | 2 | 0.48 | 0.00 |

**Πίνακας 2: Σχέσεις όπου οι ταινίες είναι αντικείμενα**

Παρατηρούμε ότι στον παραπάνω πίνακα εμφανίζονται και μη αναμενόμενες ιδιότητες, όπως για παράδειγμα οι 4 εμφανίσεις της ιδιότητας http://dbpedia.org/ontology/starring που υποδηλώνουν πως μια ταινία είναι στη θέση ηθοποιού. Πρέπει να τονίσουμε ότι τα δεδομένα εξάγονται με αυτοματοποιημένο τρόπο από τα info boxes άρθρων της Wikipedia (στα οποία υπάρχει δυνατότητα να προσθέσει πληροφορίες οποιοσδήποτε απλός χρήστης του Ιστού) και για το λόγο αυτό είναι πιθανό κάποια από αυτά να μην είναι απόλυτα αξιόπιστα. Πρόκειται όμως για τόσο μικρό ποσοστό συγκριτικά με το συνολικό όγκο πληροφορίας που φιλοξενείται στη DBpedia ώστε δεν περιμένουμε να επηρεάσει τη συνολική εικόνα.

Πέρα από τα στατιστικά που αφορούν αποκλειστικά τον αριθμό εμφάνισης των ιδιοτήτων, έχει ενδιαφέρον να εξεταστούν και οι τιμές που λαμβάνουν οι ιδιότητες αυτές, τουλάχιστον για κάποιες βασικές ιδιότητες, όπως οι ηθοποιοί και οι σκηνοθέτες. Τέτοιου είδους πληροφορίες περιλαμβάνονται στους πίνακες 3-6.

| | |
|---|---|
| Πλήθος ηθοποιών | 9666 |
| Ηθοποιοί που εμφανίζονται σε πάνω από 1 ταινίες | 3706 |
| Ηθοποιοί που εμφανίζονται σε πάνω από 5 ταινίες | 990 |
| Ηθοποιοί που εμφανίζονται σε πάνω από 10 ταινίες | 352 |
| Ηθοποιοί που εμφανίζονται σε πάνω από 15 ταινίες | 121 |
| Ηθοποιοί που εμφανίζονται σε πάνω από 20 ταινίες | 43 |

**Πίνακας 3: Στατιστικά στοιχεία ηθοποιών**

| | |
|---|---|
| Πλήθος σκηνοθετών | 2468 |
| Σκηνοθέτες που έχουν σκηνοθετήσει πάνω από 1 ταινίες | 997 |
| Σκηνοθέτες που έχουν σκηνοθετήσει πάνω από 3 ταινίες | 348 |
| Σκηνοθέτες που έχουν σκηνοθετήσει πάνω από 5 ταινίες | 167 |
| Σκηνοθέτες που έχουν σκηνοθετήσει πάνω από 10 ταινίες | 29 |
| Σκηνοθέτες που έχουν σκηνοθετήσει πάνω από 15 ταινίες | 1 |

**Πίνακας 4: Στατιστικά στοιχεία σκηνοθετών**

| | |
|---|---|
| Πλήθος σεναριογράφων | 4945 |
| Σεναριογράφοι με πάνω από 1 ταινίες | 1392 |
| Σεναριογράφοι με πάνω από 5 ταινίες | 139 |
| Σεναριογράφοι με πάνω από 10 ταινίες | 15 |

**Πίνακας 5: Στατιστικά στοιχεία σεναριογράφων**

| | |
|---|---|
| Πλήθος γλωσσών που χρησιμοποιούνται στο σύνολο των ταινιών | 133 |
| Γλώσσες που χρησιμοποιούνται σε πάνω από 5 ταινίες | 25 |
| Γλώσσες που χρησιμοποιούνται σε πάνω από 15 ταινίες | 12 |
| Γλώσσες που χρησιμοποιούνται σε πάνω από 25 ταινίες | 9 |
| Γλώσσες που χρησιμοποιούνται σε πάνω από 35 ταινίες | 6 |

**Πίνακας 6: Στατιστικά στοιχεία για τη γλώσσα των ταινιών**

Σχετικά με τα είδη των ταινιών:

- πλήθος ειδών ανά ταινία

| | |
|---|---|
| Πλήθος ταινιών που ανήκουν σε 1 είδος | 274 |
| Πλήθος ταινιών που ανήκουν σε 2 είδη | 537 |
| Πλήθος ταινιών που ανήκουν σε 3 είδη | 513 |
| Πλήθος ταινιών που ανήκουν σε 4 είδη | 303 |
| Πλήθος ταινιών που ανήκουν σε 5 είδη | 92 |
| Πλήθος ταινιών που ανήκουν σε 6 είδη | 22 |
| Πλήθος ταινιών που ανήκουν σε 7 είδη | 7 |
| Πλήθος ταινιών που ανήκουν σε 8 είδη | 2 |
| Πλήθος ταινιών που ανήκουν σε 9 είδη | 1 |
| Πλήθος ταινιών που ανήκουν σε 10 είδη | 1 |

**Πίνακας 7: Πλήθος ειδών ανά ταινία**

- πλήθος ταινιών ανά είδος

| | |
|---|---|
| Πλήθος ταινιών του είδους Crime | 340 |
| Πλήθος ταινιών του είδους HOrrOr | 159 |
| Πλήθος ταινιών του είδους ROmance | 477 |
| Πλήθος ταινιών του είδους HistOry | 49 |
| Πλήθος ταινιών του είδους Mystery | 161 |
| Πλήθος ταινιών του είδους COmedy | 647 |
| Πλήθος ταινιών του είδους SpOrt | 69 |
| Πλήθος ταινιών του είδους ShOrt | 7 |
| Πλήθος ταινιών του είδους Fantasy | 120 |
| Πλήθος ταινιών του είδους War | 98 |
| Πλήθος ταινιών του είδους ActiOn | 344 |
| Πλήθος ταινιών του είδους Adult | 12 |
| Πλήθος ταινιών του είδους Music | 61 |
| Πλήθος ταινιών του είδους Musical | 108 |
| Πλήθος ταινιών του είδους Thriller | 488 |
| Πλήθος ταινιών του είδους BiOgraphy | 59 |
| Πλήθος ταινιών του είδους AnimatiOn | 30 |
| Πλήθος ταινιών του είδους Talk-ShOw | 1 |
| Πλήθος ταινιών του είδους DOcumentary | 32 |
| Πλήθος ταινιών του είδους Western | 54 |
| Πλήθος ταινιών του είδους Adventure | 186 |
| Πλήθος ταινιών του είδους Film-NOir | 23 |
| Πλήθος ταινιών του είδους Family | 145 |
| Πλήθος ταινιών του είδους Sci-Fi | 114 |

| | |
|---|---|
| Πλήθος ταινιών του είδους Drama | 991 |

**Πίνακας 8: Πλήθος ταινιών ανά είδος**

Τέλος, παραθέτουμε στοιχεία σχετικά με τις Wikipedia κατηγορίες στις οποίες ανήκει το άρθρο που αντιστοιχεί στην εγγραφή κάθε ταινίας της DBpedia. Οι πληροφορίες αυτές είναι διαθέσιμες αφενός μέσω του λεξιλογίου SKOS και αφετέρου μέσω της οντολογίας YAGO, για τα οποία θα μιλήσουμε αναλυτικότερα σε επόμενο κεφάλαιο.

SKOS: (Τα στοιχεία αφορούν 5171 ταινίες.)

| | |
|---|---|
| Πλήθος SKOS κατηγοριών που εμφανίζονται | 4612 |
| Κατηγορίες που περιλαμβάνουν πάνω από 5 ταινίες | 1158 |
| Κατηγορίες που περιλαμβάνουν πάνω από 10 ταινίες | 718 |
| Κατηγορίες που περιλαμβάνουν πάνω από 25 ταινίες | 324 |
| Κατηγορίες που περιλαμβάνουν πάνω από 50 ταινίες | 166 |
| Κατηγορίες που περιλαμβάνουν πάνω από 80 ταινίες | 93 |
| Κατηγορίες που περιλαμβάνουν πάνω από 100 ταινίες | 63 |
| Κατηγορίες που περιλαμβάνουν πάνω από 200 ταινίες | 22 |
| Κατηγορίες που περιλαμβάνουν πάνω από 300 ταινίες | 3 |

**Πίνακας 9: Στατιστικά στοιχεία SKOS κατηγοριών**

YAGO: (Τα στοιχεία αφορούν 4335 ταινίες.)

| | |
|---|---|
| Πλήθος YAGO κλάσεων που εμφανίζονται | 2550 |
| Κλάσεις στις οποίες ανήκουν πάνω από 5 ταινίες | 732 |
| Κλάσεις στις οποίες ανήκουν πάνω από 10 ταινίες | 482 |
| Κλάσεις στις οποίες ανήκουν πάνω από 20 ταινίες | 279 |
| Κλάσεις στις οποίες ανήκουν πάνω από 50 ταινίες | 112 |
| Κλάσεις στις οποίες ανήκουν πάνω από 80 ταινίες | 67 |
| Κλάσεις στις οποίες ανήκουν πάνω από 100 ταινίες | 47 |
| Κλάσεις στις οποίες ανήκουν πάνω από 150 ταινίες | 24 |
| Κλάσεις στις οποίες ανήκουν πάνω από 200 ταινίες | 6 |

**Πίνακας 10: Στατιστικά στοιχεία YAGO κλάσεων**

Από τα στοιχεία που παρουσιάστηκαν είναι προφανές ότι οι πληροφορίες που συγκεντρώσαμε για τις 5179 ταινίες που προέκυψαν από την αντιστοίχιση των ταινιών του Netflix στις ταινίες της DBpedia συνθέτουν ένα πολύ πλούσιο σύνολο δεδομένων. Το σημαντικότερο συμπέρασμα που προκύπτει από τα στατιστικά στοιχεία που προηγήθηκαν είναι πως οι ταινίες μας μπορούν να ομαδοποιηθούν με κριτήριο πολλά από τα χαρακτηριστικά τους (ταινίες του ίδιου είδους, ταινίες του ίδιου σκηνοθέτη, ταινίες της ίδιας YAGO κλάσης κ.α.). Για το λόγο αυτό αυξάνονται οι πιθανότητες να ανακαλύψουμε πρότυπα που περιγράφουν τις επιλογές ταινιών των χρηστών.

## *3.3 Κριτήρια επιλογής χρηστών*

Έχοντας εξασφαλίσει ένα πλούσιο σύνολο πληροφοριών σχετικά με επιλεγμένες ταινίες που έχουν βαθμολογήσει οι χρήστες του Netflix, είναι σημαντικό να ξεχωρίσουμε από το σύνολο των χρηστών τους καταλληλότερους για τους σκοπούς της εφαρμογής μας. Υπενθυμίζουμε πως για να προχωρήσουμε στην εξατομίκευση της αναζήτησης ενός χρήστη, προαπαιτούμενη είναι η δημιουργία του προφίλ των ενδιαφερόντων του. Σκοπός μας είναι, από τη βαθμολογία κάθε ταινίας, αξιοποιώντας τις πληροφορίες που τη συνοδεύουν, να ανακαλύψουμε τα πρότυπα που ακολουθεί ο χρήστης, εφόσον αυτά υπάρχουν, για να αποφασίσει τις βαθμολογίες που δίνει. Για να αυξήσουμε τις πιθανότητες να επιλεχθούν χρήστες πιο πρόσφοροι για την ανακάλυψη τέτοιων προτύπων συμπεριφοράς, χρησιμοποιούμε τα εξής κριτήρια:

α) Πλήθος βαθμολογιών

Για να εκπαιδευτεί επιτυχώς το SVM, αλλά και για να διασφαλίσουμε αρκετή ευελιξία στα πειράματά μας, θα πρέπει να έχουμε στη διάθεσή μας επαρκή δεδομένα.

β) Ποσοστό βαθμολογιών που αφορούν τις 5179 ταινίες/ αρχικός αριθμός βαθμολογιών

Είναι σημαντικό οι βαθμολογίες που έχουν απομείνει για κάθε χρήστη να αποτελούν ένα, όσο γίνεται, πιο αντιπροσωπευτικό δείγμα των επιλογών του.

γ) Ποσοστό καλών-κακών βαθμολογιών

Δεδομένου ότι το πλήθος των διαφορετικών βαθμολογιών είναι ήδη πολύ μικρό (ακέραιες τιμές στο διάστημα [1,5] ), θέλουμε να εξασφαλίσουμε τη μεγαλύτερη δυνατή ποικιλία στις βαθμολογίες του χρήστη. Ένας χρήστης που είναι μόνιμα πολύ αυστηρός ή πολύ επιεικής, άρα έχει δώσει την ίδια βαθμολογία στην πλειοψηφία των ταινιών του, δεν προσφέρεται για τους σκοπούς των πειραμάτων πάνω στην εξατομίκευση αναζήτησης.

# 4

# *Εκμετάλλευση σημασιολογικής πληροφορίας και επιλογή μεθόδου εξατομίκευσης*

Σε αυτό το κεφάλαιο παρουσιάζεται η εξαγωγή σημασιολογικών πληροφοριών από διαθέσιμες οντολογίες και άλλες ιεραρχικές δομές που περιγράφουν τις σχέσεις μεταξύ των εγγραφών της DBpedia που μας ενδιαφέρουν. Επιπλέον εξηγούνται οι λόγοι για τους οποίους απορρίφθηκε η χρήση μιας εκ των δύο μεθόδων εξατομίκευσης που παρουσιάστηκαν σε προηγούμενο κεφάλαιο.

## *4.1 Καθορισμός ιεραρχικών σχέσεων rdf δεδομένων*

Γνωρίζουμε πως η ύπαρξη σχέσεων ιεραρχίας μεταξύ των δεδομένων του rdf γράφου μπορεί να αξιοποιηθεί για την αποτελεσματικότερη αναζήτηση σε αυτά. Επιπλέον, είδαμε στο κεφάλαιο 2 πως πολλά χαρακτηριστικά εκπαίδευσης που χρησιμοποιούνται για το Ranking SVM εξαρτώνται από την ιεραρχία κλάσεων του υποκείμενου μοντέλου δεδομένων του rdf γράφου. Σε αυτή την ενότητα θα ερευνήσουμε την ύπαρξη τέτοιων πληροφοριών που να μπορούν να αξιοποιηθούν για τα δικά μας δεδομένα.

### *4.1.1 Οντολογία*

"Μια οντολογία είναι μια ρητή και τυπική προδιαγραφή μιας επίνοιας (conceptualization)", [Studer 1998]. Οι οντολογίες παρέχουν μια κοινή κατανόηση ενός πεδίου, μέσω του ορισμού των σημαντικών εννοιών (κλάσεων) του πεδίου και τον καθορισμό των μεταξύ τους σχέσεων. Με τη βοήθεια των οντολογιών, οι μηχανές αναζήτησης μπορούν να βελτιώσουν την ακρίβεια των αναζητήσεων αξιοποιώντας τις εννοιολογικές πληροφορίες του υποκείμενου μοντέλου και της ιεραρχίας κλάσεων. Για τον ίδιο λόγο, εξέχουσα είναι η θέση των οντολογιών και στον τομέα της εξατομίκευσης της αναζήτησης. Δεν είναι τυχαίο ότι πολλά από τα χαρακτηριστικά εκπαίδευσης των Μηχανών Διανυσμάτων Υποστήριξης (SVM) που αναφέρθηκαν στο κεφάλαιο 2 προϋποθέτουν την ύπαρξη μιας οντολογίας για την εννοιολογική αναπαράσταση των δεδομένων στα οποία πραγματοποιείται η αναζήτηση.

Για την ανάπτυξη μιας οντολογίας είναι σημαντικός ο προσδιορισμός της σκοπιμότητας και του πεδίου εφαρμογής της, δηλαδή ποιοι θα είναι οι χρήστες της και γιατί θα τη χρησιμοποιήσουν. Για τα δεδομένα της DBpedia, η χρησιμοποιούμενη οντολογία γραμμένη στη γλώσσα OWL, ορίζει την εξής ιεραρχία για την κλάση Film:

owl: Thing
    Work
        Film

Η κλάση owl:Thing βρίσκεται στην κορυφή της ιεραρχίας, ενώ η Film δεν έχει υποκλάσεις.

Για την κλάση Actor:

owl: Thing
    Agent
        Person
            Artist
                Actor
                    AdultActor
                    VoiceActor

Για τους σκηνοθέτες δεν υπάρχει ειδική κλάση, οπότε χρησιμοποιείται η ευρύτερη κλάση Person.

Βλέπουμε πως όλες οι ταινίες ανήκουν στην ίδια κλάση και βρίσκονται συνεπώς στο ίδιο επίπεδο της ιεραρχίας κλάσεων. Δεδομένου ότι ο χώρος της αναζήτησης για τους σκοπούς της έρευνάς μας περιλαμβάνει κατά κύριο λόγο αντικείμενα των τριών αυτών κατηγοριών, η παραπάνω οντολογία δεν προσφέρει αξιοποιήσιμες πληροφορίες. Είναι φυσικά χρήσιμη για το στάδιο ανάκτησης αποτελεσμάτων από τη μηχανή αναζήτησης, αλλά δεν παρέχει δυνατότητα αναζήτησης ταινιών με πιο συγκεκριμένα κριτήρια όπως το είδος, τη θεματολογία τους, τα βραβεία που έχουν λάβει. Επιπλέον, για τους ίδιους λόγους, δεν μπορεί

να βοηθήσει στην εξατομίκευση της αναζήτησης. Για το σκοπό αυτό χρήσιμη θα ήταν μια οντολογία που παρέχει μια ιεραρχία υποκλάσεων της κλάσης Film (αντίστοιχα για ηθοποιούς και σκηνοθέτες) .

Τέτοιου είδους πληροφορίες μπορούν να εξαχθούν από τις κατηγορίες στις οποίες κατατάσσονται τα άρθρα της Wikipedia, αφού ξέρουμε πως κάθε εγγραφή της DBpedia αντιστοιχεί σε ένα άρθρο της Wikipedia.

### *4.1.2 Εξαγωγή σχέσεων ιεραρχίας από τις κατηγορίες άρθρων της Wikipedia*

Οι κατηγορίες άρθρων στη Wikipedia σχηματίζουν μια πολυ-ιεραρχική δομή όπου η σχέση κατηγορίας-υποκατηγορίας είναι πολύ πιο χαλαρή από τη σχέση κλάσης-υποκλάσης (http://en.wikipedia.org/wiki/Wikipedia:Categorization). Για την αναπαράσταση αυτών των πληροφοριών σε RDF μορφή χρησιμοποιείται το λεξιλόγιο SKOS (http:// www.w3.org/TR/skos-reference/). Κάθε κατηγορία μπορεί να έχει παραπάνω από μια υπερ-κατηγορίες οι οποίες μπορούν να θεωρηθούν εννοιολογικά ευρύτερες αυτής. Για τον ορισμό των κατηγοριών στις οποίες ανήκει μια εγγραφή της DBpedia χρησιμοποιείται η σχέση http://purl.org/dc/terms/subject ενώ για να δηλωθεί ότι η κατηγορία A είναι ευρύτερη της B χρησιμοποιείται η τριπλέτα

<B> <http://www.w3.org/2004/02/skos/core#broader> <A>

Ξεκινώντας από την ταινία Million Dollar Baby και ακολουθώντας την ιεραρχία των κατηγοριών στις οποίες ανήκει συναντάται το εξής μονοπάτι:

< Million Dollar Baby > <subject> < Category: American sports films >

< Category: American sports films > <broader> <Category: Sports media in the United States>

<Category: Sports media in the United States> <broader> <Category: Sports in the United States>

<Category: Sports in the United States> <broader> <Category: American culture>

<Category: American culture> <broader> <Category: American studies>

<Category: Category: American studies> <broader> <Category: American culture>

όπου το <broader> αντιπροσωπεύει τη σχέση <http://www.w3.org/2004/02/skos/core#broader>

και το <subject> τη σχέση <http://purl.org/dc/terms/subject>, ενώ τα URI που εμφανίζονται ως υποκείμενα και αντικείμενα έχουν συντομευθεί για τη διευκόλυνση της παρουσίασής τους.

Τυπικά δεν πρόκειται για μονοπάτι εξαιτίας του κύκλου που σχηματίζουν οι δύο τελευταίες ακμές. Καταδεικνύονται μέσω του παραδείγματος δύο πολύ σημαντικά προβλήματα.

Πρώτον, η έλλειψη αυστηρών σημασιολογικών απαιτήσεων για τον ορισμό μιας κατηγορίας ως ευρύτερη μιας άλλης, επιφέρει μια ασάφεια στο τι αντιπροσωπεύει τελικά η σχέση <Broader>. Στο παράδειγμά μας, δεν είναι προφανές πως από τις αμερικανικές ταινίες που

σχετίζονται με σπορ καταλήξαμε στις Αμερικανικές Σπουδές. Δεύτερον, η ύπαρξη δύο κατηγοριών που αποτελούν ταυτόχρονα υποκατηγορία και υπερκατηγορία η μια της άλλης είναι μάλλον παράδοξη.

Η συμπεριφορά αυτή της ιδιότητας <Broader> δεν είναι ίσως διαισθητικά προφανής, είναι όμως απόλυτα συμβατή με τους κανόνες του μοντέλου/λεξιλογίου SKOS, σύμφωνα με τους οποίους η ιδιότητα αυτή δε δηλώνεται ως μεταβατική ούτε ως μη ανακλαστική. Η παρεχόμενη ευελιξία δίνει πρόσφορο έδαφος για την αναπαράσταση της πολύπλοκης δομής κατηγοριών των άρθρων της Wikipedia. Είναι, συνεπώς, αναγκαίος ο καθορισμός ενός άλλου, αυστηρότερου πλαισίου εξαγωγής της διαθέσιμης πληροφορίας από τις κατηγορίες της Wikipedia.

#### *4.1.2.1 Εξαγωγή πληροφοριών από τις SKOS κατηγορίες*

Στο αρχείο Articles Categories της DBpedia περιέχονται για κάθε οντότητα οι κατηγορίες στις οποίες ανήκει και στο αρχείο Categories (SKOS) οι σχέσεις μεταξύ κατηγοριών. Από το δεύτερο αρχείο μας ενδιαφέρουν μόνο οι τριπλέτες που αφορούν τη σχέση <Broader>.

Ζητούμενο είναι να χρησιμοποιήσουμε τις πληροφορίες αυτές ώστε να εμπλουτίσουμε τη βάση δεδομένων μας με νέα δεδομένα και να ανακαλύψουμε συσχετίσεις μεταξύ ταινιών/ηθοποιών/συγγραφέων που δεν είναι άμεσα διαθέσιμες, αποφεύγοντας τα προβλήματα που αναφέρθηκαν παραπάνω. Δεν επιδιώκουμε την ανάπτυξη μιας οντολογίας που θα οργανώσει τις κατηγορίες σε μορφή δέντρου. Πιο συγκεκριμένα, δεν αποτελεί πρόβλημα η αντιστοίχιση κάθε ταινίας σε παραπάνω από μια κατηγορίες ούτε υπάρχει απαίτηση κάθε κατηγορία να έχει μια μόνο υπερ-κατηγορία, ακόμα και αν αυτό εκ των πραγμάτων αποκλείει τη δυνατότητα σχηματισμού αυστηρής ιεραρχίας. Δεδομένου άλλωστε ότι το λεξιλόγιο SKOS επιτρέπει τη χρήση της σχέσης broader για τη δημιουργία σχέσεων ιεραρχίας της μορφής

A broader B
B broader C
A broader C

ο καθορισμός της θέσης της κατηγορίας Α στην ιεραρχία των κατηγοριών δεν είναι καθόλου εύκολος. Στόχος μας είναι ουσιαστικά να περιορίσουμε το πλήθος των κατηγοριών που συσχετίζονται με ταινίες και να κρατήσουμε μόνο τη χρήσιμη πληροφορία ώστε να αναδειχθούν οι κατηγορίες που μπορούν να χρησιμεύσουν ως χαρακτηριστικά στο SVM (SVM features).

Οι κατηγορίες στις οποίες ανήκει μια ταινία μπορεί να παρέχουν στοιχεία, μεταξύ άλλων, για το είδος της, την εταιρεία παραγωγής της, το έτος κυκλοφορίας της, το σκηνοθέτη της, το

αντικείμενό της (π.χ. Πρώτος Παγκόσμιος Πόλεμος, Αυτοκτονία και Κατάθλιψη, Ιππασία) και την τεχνική κινηματογράφησής της. Η πλειοψηφία των κατηγοριών στις οποίες ανήκει άμεσα μια ταινία φέρουν χρήσιμη πληροφορία. Χωρίς την αξιοποίηση, όμως, των ιεραρχικών σχέσεων που συνδέουν τις κατηγορίες μεταξύ τους, το σύστημα εξατομίκευσης δεν μπορεί να κατανοήσει ότι μια ταινία της κατηγορίας "Teen-Comedy" συνδέεται περισσότερο με μια ταινία της κατηγορίας "Comedy" συγκριτικά με μια ταινία της κατηγορίας "Science-Fiction". Από την άλλη πλευρά, η αλόγιστη μετάβαση από μια κατηγορία στις ευρύτερές της μπορεί για τους λόγους που είδαμε να συγκεντρώσει παραπλανητικά μεγάλο αριθμό ταινιών σε κάποια μη σχετική κατηγορία.

Ξεκινώντας από τις κατηγορίες που δεν είναι ευρύτερες (broader) καμίας άλλης (μηδενικού επιπέδου κατηγορίες), ορίσαμε μια ιεραρχία με κριτήριο τον αριθμό των σχέσεων <Broader> που χρειάστηκαν ώστε να φτάσουμε από τις κατηγορίες του μηδενικού επιπέδου σε κάποια άλλη κατηγορία. Ξεκινώντας από τις κατηγορίες κάθε ταινίας και ακολουθώντας όλα τα μονοπάτια βρήκαμε για κάθε ταινία σε ποιες κατηγορίες μπορούμε να φτάσουμε. Ωστόσο, για τους λόγους που εξηγήσαμε παραπάνω, μεταχειριστήκαμε όλες τις κατηγορίες με τον ίδιο τρόπο, χωρίς να μας ενδιαφέρει αν μια ταινία ανήκει άμεσα ή έμμεσα (και μετά από πόσα βήματα) σε καθεμία. Στη συνέχεια, αποκλείσαμε τις κατηγορίες των οποίων ο τίτλος δε συμφωνεί με τα πρότυπα που ορίσαμε ότι δείχνουν συσχέτιση της κατηγορίας με το χώρο του κινηματογράφου.

Παρόμοια λογική ακολουθούμε και για την εύρεση όλων των κατηγοριών στις οποίες ανήκουν οι ηθοποιοί και οι σκηνοθέτες των ταινιών μας.

#### *4.1.2.2 Χρήση της οντολογίας YAGO*

Η ιεραρχία κλάσεων της οντολογίας (http://www.mpi-inf.mpg.de/yago-naga/yago/) είναι διαθέσιμη από τη DBpedia μαζί με την αντιστοίχιση κάθε οντότητας στην κλάση που ανήκει. Για την ανάπτυξη της οντολογίας YAGO έχουν χρησιμοποιηθεί τεχνικές που συνδέουν την ιεραρχία κατηγοριών της Wikipedia με την ιεραρχία εννοιών του WordNet [SKW07]. Με τον τρόπο αυτό, από μια ακατάλληλη για οντολογικούς σκοπούς ιεραρχία, αξιοποιώντας την ταξονομία εννοιών του WordNet, κατέστη δυνατός ο ορισμός κλάσεων και των μεταξύ τους σχέσεων για τις εκατομμύρια οντότητες που περιλαμβάνει η Wikipedia (και συνεπώς η DBpedia). Αποδεικνύεται ότι οι υπερκλάσεις των κλάσεων στις οποίες ανήκουν οι ταινίες μας δεν προσφέρουν κάποια επιπλέον πληροφορία, γεγονός λογικό αφού, τελικά, πρόκειται πράγματι για οντότητες του ίδιου επιπέδου. Κρατάμε όμως τις κλάσεις στις οποίες ανήκουν άμεσα οι ταινίες και θα τις χρησιμοποιήσουμε όπως τις SKOS κατηγορίες που περιγράψαμε στο 4.1.2.1 ώστε να επιλέξουμε τελικά όποια από τις δύο προσεγγίσεις έχει καλύτερα

αποτελέσματα στα πειράματά μας στην εξατομίκευση της αναζήτησης. Αντίστοιχα, κρατάμε τις κλάσεις στις οποίες ανήκουν οι ηθοποιοί και οι σκηνοθέτες του συνόλου δεδομένων μας.

## *4.2 Επιλογή κατάλληλου συστήματος εξατομίκευσης*

Στο κεφάλαιο 2 περιγράφηκαν δύο μέθοδοι εξατομίκευσης αναζήτησης σε rdf δεδομένα και αναφέρθηκε πως στην παρούσα εργασία θα χρησιμοποιηθεί μόνο η μέθοδος του Ranking SVM. Στην ενότητα αυτή θα εξηγήσουμε γιατί η μέθοδος του propagation δεν ενδείκνυται για τους σκοπούς της εργασίας μας, λαμβάνοντας υπόψη τις ιδιαιτερότητες των δεδομένων που έχουμε στη διάθεσή μας.

### *4.2.1 Χρήση propagation*

Από το κεφάλαιο 3 γνωρίζουμε ότι τα αρχικά μας δεδομένα περιλαμβάνουν βαθμολογίες ταινιών οι οποίες μπορούν να χρησιμοποιηθούν ως αρχικά βάρη στους κόμβους του γράφου που αντιπροσωπεύουν αυτές τις ταινίες. Προφανώς δε διαθέτουμε πληροφορίες για την ανάθεση βαρών σε ακμές του γράφου, δε γνωρίζουμε δηλαδή ποιες σχέσεις ενδιαφέρουν περισσότερο τους χρήστες.

Καθώς, όπως εξηγήσαμε στην παράγραφο 4.1 δε διαθέτουμε μια οντολογία που να μπορεί να χρησιμοποιηθεί για τον ορισμό ιεραρχίας κλάσεων ταινιών, η εξάπλωση δεν μπορεί να γίνει μέσω ακμών κλάσης-υποκλάσης, οπότε για το σκοπό αυτό θα πρέπει να επιλεγούν άλλου είδους ακμές. Η πιο κοντινή στην έννοια της κλάσης πληροφορία που διαθέτουμε για τις ταινίες είναι οι SKOS κατηγορίες στις οποίες ανήκουν. Ωστόσο, οι κατηγορίες SKOS αφορούν πολύ διαφορετικές ιδιότητες των ταινιών όπως θεματολογία (π.χ. Aviation_Films), είδος (π.χ. Drama_Films), σκηνοθέτη (Films_directed_by...), βραβεία, χρονολογίες και άλλα. Θα ήταν συνεπώς άστοχο να μεταχειριστούμε όλες αυτές τις κατηγορίες με τον ίδιο τρόπο χωρίς επιπλέον πληροφορίες για το χρήστη. Θα μπορούσε για το σκοπό αυτό να οριστεί κατάλληλος συντελεστής ανάλογα με το πλήθος των οντοτήτων που ανήκουν σε μια κατηγορία, όπως είδαμε σε άλλες εργασίες στο κεφάλαιο 2. Επιπρόσθετα, θα μπορούσαμε να μειώνουμε το συντελεστή εξάπλωσης καθώς μέσω των ακμών <broader> απομακρυνόμαστε από τον αρχικό κόμβο και τις κατηγορίες στις οποίες ανήκει άμεσα. Ωστόσο, εξαιτίας των ιδιαιτεροτήτων της ιεραρχίας των SKOS κατηγοριών που περιγράφηκαν σε προηγούμενη παράγραφο, ο αριθμός αυτός δεν μπορεί να θεωρηθεί αξιόπιστο κριτήριο για τον ορισμό κατάλληλου συντελεστή εξάπλωσης.

Στα πλαίσια του δικού μας χώρου αναζήτησης, θα ήταν λογικό να υποθέσουμε πως από τη βαθμολογία μίας ταινίας μπορούν να εξαχθούν συμπεράσματα για τους συντελεστές της, τουλάχιστον για τους ηθοποιούς και το σκηνοθέτη. Θα μπορούσαν συνεπώς να

χρησιμοποιηθούν αυτές οι ακμές για την εξάπλωση βαρών. Δεδομένου ότι δε γνωρίζουμε αν κάποιος από αυτούς τους δύο παράγοντες είναι πιο σημαντικός για το χρήστη, έχουμε δύο επιλογές. Μπορούμε να χρησιμοποιήσουμε τον ίδιο συντελεστή εξάπλωσης και για τους δύο αυτούς τύπους σχέσεων/ακμών. Εναλλακτικά μπορούμε να δώσουμε μεγαλύτερο βάρος στο σκηνοθέτη, αφού μια ταινία έχει πολλούς ηθοποιούς, αλλά λιγότερους σκηνοθέτες, συνηθέστερα μάλιστα μόνο έναν. Η επιλογή σε αυτό το σημείο είναι μάλλον αυθαίρετη.

Επιπλέον, ακόμα και αν χρησιμοποιήσουμε μια πολύ απλή προσέγγιση στην εξάπλωση βαρών, χρησιμοποιώντας ίδιους συντελεστές και προχωρώντας σε μονοπάτια το πολύ μήκους 2, ώστε να αποφύγουμε την παραπλανητική εξάπλωση βαρών μακριά από τα ενδιαφέροντα του χρήστη, η μέθοδος του propagation δεν μπορεί να χρησιμοποιηθεί στην εύρεση συσχετίσεων μεταξύ των διαφόρων χαρακτηριστικών μιας ταινίας. Μετά την ολοκλήρωση της εξάπλωσης των βαρών κάποιοι κόμβοι θα έχουν συγκεντρώσει μεγαλύτερα βάρη. Ωστόσο, δεν υπάρχει τρόπος να καταλάβουμε αν, για παράδειγμα, ένας ηθοποιός απέκτησε μεγαλύτερη βαθμολογία μέσω των κωμικών ταινιών του ή των δραματικών. Τέτοιου είδους συσχετίσεις μας είναι χρήσιμες για την αξιολόγηση μιας νέας ταινίας του.

Συμπεραίνουμε ότι η μέθοδος του propagation δεν μπορεί αυτόνομα να χρησιμοποιηθεί ως μέθοδος εξατομίκευσης αναζήτησης εξαιτίας του περιορισμένου συνόλου δεδομένων που έχουμε στη διάθεσή μας και των προβλημάτων που προκύπτουν.

Εξετάζουμε λοιπόν το συνδυασμό των δύο μεθόδων. Συγκεκριμένα την εφαρμογή, σε πρώτο στάδιο, της απλούστερης δυνατής (δηλαδή με τον ίδιο συντελεστή) εξάπλωσης βαρών από τις ταινίες στους ηθοποιούς και τους σκηνοθέτες ώστε να εμπλουτίσουμε τα δεδομένα εισόδου. Σε δεύτερο στάδιο, το εμπλουτισμένο με βαθμολογίες ηθοποιών και σκηνοθετών σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για την εκπαίδευση του Ranking SVM.

Ωστόσο, η ιδέα τελικά εγκαταλείφθηκε για δύο λόγους. Πρώτον, δεν είναι σαφές αν η βαθμολογία που θα αποκτήσει ένας συντελεστής μιας ταινίας, είτε με τη μορφή μέσου όρου είτε με τη μορφή αθροίσματος, είναι πράγματι αντιπροσωπευτική της γνώμης του χρήστη για εκείνον. Συνεπώς η χρήση αυτής της πληροφορίας στην εκπαίδευση του SVM είναι πιθανό να έχει ανεπιθύμητα αποτελέσματα. Δεύτερον, εφόσον οι βαθμολογίες αυτές προέρχονται από τις βαθμολογίες των ταινιών οι οποίες περιλαμβάνονται στα δεδομένα εκπαίδευσης του SVM, οι αντίστοιχες πληροφορίες είναι ήδη έμμεσα διαθέσιμες στο σύστημα και άρα η επανάληψή τους δεν προσφέρει κανένα όφελος.

Για το λόγο αυτό, εγκαταλείφθηκε οριστικά η χρήση της μεθόδου στα πλαίσια της παρούσας εργασίας.

### *4.2.2 Χρήση Ranking SVM*

Το Ranking SVM είναι η μέθοδος εξατομίκευσης αναζήτησης που θα χρησιμοποιήσουμε στην παρούσα εργασία και για το λόγο αυτό θα περιγράψουμε εκτενώς τον τρόπο λειτουργίας του γενικά, αλλά και τις ιδιαιτερότητες της δικής μας προσέγγισης σε ξεχωριστό κεφάλαιο.

# 5

# *Εκπαίδευση συναρτήσεων αναταξινόμησης με βάση σημασιολογική και κειμενική πληροφορία*

Στο κεφάλαιο αυτό γίνεται η παρουσίαση της εκπαίδευσης του συστήματος εξατομικευμένης αναταξινόμησης αποτελεσμάτων αναζήτησης. Εξηγείται ο τρόπος λειτουργίας του Ranking SVM αλλά και οι περιορισμοί που εισάγονται λόγω της έλλειψης κατάλληλων δεδομένων εισόδου. Παρουσιάζεται στη συνέχεια αναλυτικά η δημιουργία χαρακτηριστικών εκπαίδευσης που να ενσωματώνουν το σύνολο της διαθέσιμης πληροφορίας.

## *5.1 Δεδομένα εκπαίδευσης*

Ο αλγόριθμος που χρησιμοποιείται από το Ranking SVM προορίζεται για χρήση σε αναδιάταξη αποτελεσμάτων αναζήτησης ώστε η τελική κατάταξη των αποτελεσμάτων να είναι αντιπροσωπευτική των προτιμήσεων του χρήστη. Ως δεδομένα εισόδου για την εκπαίδευση του SVM χρησιμοποιούνται ερωτήματα με τις ήδη ταξινομημένες λίστες αποτελεσμάτων τους. Κατά την εκπαίδευση, στα αποτελέσματα που αφορούν το ίδιο ερώτημα ανατίθεται ο ίδιος αναγνωριστικός αριθμός. Μετά την εκπαίδευση, το σύστημα μπορεί να αναδιατάξει λίστες νέων αποτελεσμάτων τα οποία πρέπει να ανταποκρίνονται σε κάποιο από τα ερωτήματα που χρησιμοποιήθηκαν στην εκπαίδευση, έχουν δηλαδή τον ίδιο

αναγνωριστικό αριθμό με κάποιο από αυτά. Η χρήση πολλών ερωτημάτων αναδεικνύει τη διαφορετικότητα μεταξύ των αναζητήσεων ενός χρήστη και την ανάγκη αναταξινόμησης αποτελεσμάτων μέσα από το πρίσμα του ερωτήματος που οδήγησε στην ανάκτησή τους. Ειδικότερα, καθίσταται έτσι δυνατό να αξιοποιηθούν τα ιδιαίτερα χαρακτηριστικά κάθε ερωτήματος, όπως το πόσο εξειδικευμένες μπορεί να είναι οι πληροφορίες που αναζητά ο χρήστης σε κάθε περίπτωση ή σε ποιο γνωστικό πεδίο εντάσσονται. Υπενθυμίζεται ξανά ότι το σύστημα χρησιμοποιείται μόνο για τη σύγκριση αποτελεσμάτων μεταξύ τους και πως ο βαθμός που ανατίθεται εσωτερικά σε καθένα από αυτά δεν έχει αξία έξω από αυτό το πλαίσιο.

Έχουμε σε προηγούμενο κεφάλαιο αναφερθεί στην έλλειψη δεδομένων εισόδου αντίστοιχων με αυτά που περιγράφηκαν παραπάνω. Είναι σημαντικό να γίνει τώρα μια αναλυτική παρουσίαση της διαφορετικής μορφής των δεδομένων μας, των περιορισμών που αυτή επιφέρει, καθώς και της προσέγγισης που θα ακολουθήσουμε ώστε να πετύχουμε το στόχο μας.

Διαθέτουμε μια μεγάλη λίστα αποτελεσμάτων τα οποία αντιστοιχούν, ουσιαστικά, στο ίδιο ερώτημα. Οι ταινίες με τις βαθμολογίες κάθε χρήστη μπορούν να θεωρηθούν αποτέλεσμα ενός ερωτήματος με μόνη λέξη κλειδί τη λέξη "Ταινία". Αυτό φυσικά αφορά την εκπαίδευση του συστήματος. Στις αναζητήσεις του χρήστη μπορούμε να διαχειριστούμε και φράσεις-κλειδιά, όπως "drama film" ή "woody allen film", τα αποτελέσματα των οποίων θα είναι υποσύνολο των συνολικών ταινιών κάθε χρήστη. Οι πιθανές βαθμολογίες είναι ακέραιοι αριθμοί στο διάστημα [1,5], γεγονός που περιορίζει σημαντικά τα εξαγόμενα συμπεράσματα για τις προτιμήσεις του χρήστη. Στην πραγματικότητα, σε μια λίστα αποτελεσμάτων που ανακτήθηκαν για ένα συγκεκριμένο ερώτημα, περιμένουμε ο χρήστης να έχει για πολλά από αυτά αρκετά διαφορετική αντίληψη ως προς τη σχετικότητά τους με τις πληροφορίες που αναζητά. Η αυστηρή διάταξη των αποτελεσμάτων δεν είναι απαίτηση του αλγόριθμου, ούτε και θα ήταν φυσικά εφικτό ο χρήστης να δίνει σε κάθε αποτέλεσμα διαφορετική βαθμολογία. Είναι όμως σαφώς πιο περιορισμένο το πλήθος των αποτελεσμάτων που θα θεωρούνταν εξίσου σχετικά με το ερώτημα που τέθηκε, τουλάχιστον σε ό,τι αφορά τα αποτελέσματα που βρίσκονται ψηλότερα στη λίστα. Το γεγονός αυτό, σε συνδυασμό με το πολύ μεγαλύτερο μέγεθος της λίστας συγκριτικά με τα αποτελέσματα που αναμένουμε να εξετάσει ο χρήστης σε μια μηχανή αναζήτησης, αναμένουμε να επηρεάσει τη συμπεριφορά του συστήματος.

Εξετάστηκε η δυνατότητα διαχωρισμού της λίστας ταινιών που θα χρησιμοποιηθούν στην εκπαίδευση του SVM σε πολλές επιμέρους λίστες που θα αντιστοιχίζονται σε διαφορετικά ερωτήματα. Η διαίσθηση πίσω από μια τέτοια επιλογή είναι η αποτύπωση της χρονικής εξέλιξης των προτιμήσεων ενός χρήστη. Ωστόσο, οι ημερομηνίες που ένας χρήστης βαθμολόγησε τις διάφορες ταινίες, κρύβουν πολύ σημαντικότερη και πολύ διαφορετική

συχνά πληροφορία από το πότε είδε καθεμία από αυτές.[PC09] Θα ήταν συνεπώς άστοχος και πιθανώς παραπλανητικός ο διαχωρισμός της συνολικής λίστας ταινιών σε πολλά υποσύνολα με απλά ημερολογιακά κριτήρια.

Εναλλακτικός τρόπος διαχωρισμού των ταινιών του συνόλου εκπαίδευσης, θα ήταν με κριτήριο το είδος τους. Στόχος μας είναι, όμως, το SVM να εκπαιδευτεί να συγκρίνει ταινίες και διαφορετικού είδους. Εφόσον αυτός ο στόχος είναι πιο γενικός και παρουσιάζει μεγαλύτερο ενδιαφέρον, κρίνεται σκόπιμο να μην ακολουθηθεί μια τέτοια προσέγγιση. Η αντίστοιχη παρατήρηση ισχύει και για διαχωρισμό με κριτήριο κάποιο άλλο χαρακτηριστικό των ταινιών, όπως η χώρα παραγωγής ή η γλώσσα.

Ως αποτέλεσμα της μη ύπαρξης πολλών διακριτών ερωτημάτων και ταυτόχρονα της ύπαρξης δεδομένων εισόδου αποκλειστικά ενός είδους (ταινίες), είναι δυνατή μόνο η δημιουργία χαρακτηριστικών εκπαίδευσης ανεξάρτητων από το ερώτημα του χρήστη. Το σύστημα δηλαδή θα ταξινομεί με τον ίδιο τρόπο όμοια αποτελέσματα που έχουν ανακτηθεί σε διαφορετικές αναζητήσεις του χρήστη.

## *5.2 Είδος αποτελεσμάτων προς εξατομικευμένη ταξινόμηση*

Ενώ στα δεδομένα εισόδου περιλαμβάνονται αποκλειστικά ταινίες, στόχος είναι το SVM να μάθει να ταξινομεί λίστες που περιλαμβάνουν και άλλου είδους αποτελέσματα, σχετικά φυσικά με ταινίες. Αξιολογώντας τις διαθέσιμες πληροφορίες , καθώς και τους παράγοντες που αναμένεται να επηρεάζουν και να επηρεάζονται από τις κινηματογραφικές προτιμήσεις ενός χρήστη, κρίνεται σκόπιμη και εφικτή η επέκταση της λειτουργίας του SVM ώστε να ταξινομεί επιπλέον των ίδιων των ταινιών, τους σκηνοθέτες και τους ηθοποιούς τους.

Παρόλο που αναμένεται το SVM μετά την εκπαίδευση να έχει εξάγει πληροφορίες σχετικά με τα είδη των ταινιών που είναι περισσότερο αρεστά στο χρήστη, η πληροφορία αυτή θα αξιοποιηθεί έμμεσα για την αξιολόγηση των τριών ειδών αποτελεσμάτων που αναφέραμε. Δεδομένου ότι το είδος των ταινιών δεν περιμένουμε να εμφανιστεί ως αυτόνομο αποτέλεσμα σε κάποιο ερώτημα, παρά μόνο ίσως ως συνοδευτική πληροφορία, δεν έχει νόημα να ταξινομεί το SVM και είδη ταινιών. Το αντίστοιχο ισχύει και για τη γλώσσα μιας ταινίας ή τη χώρα στην οποία γυρίστηκε. Από την άλλη πλευρά, προς αποφυγή απλουστευτικής επέκτασης των προτιμήσεων του χρήστη, δεν υπάρχει απαίτηση εξαγωγής συμπερασμάτων για το σύνολο των χαρακτηριστικών μιας ταινίας. Οι πληροφορίες, για παράδειγμα, σχετικά με σεναριογράφους, παραγωγούς, κινηματογραφιστές, δεν κρίνονται επαρκείς για την αυτόνομη αξιολόγησή τους.

## *5.3 Χαρακτηριστικά εκπαίδευσης SVM*

Το σημαντικότερο κομμάτι στη λειτουργία του Ranking SVM είναι η επιλογή κατάλληλων χαρακτηριστικών που να εσωκλείουν όλη την απαραίτητη πληροφορία για την αξιολόγηση και εξατομικευμένη ταξινόμηση των αποτελεσμάτων ενός ερωτήματος. Κάθε αποτέλεσμα στη λίστα αντιστοιχίζεται σε ένα διάνυσμα στο N-διάστατο χώρο, όπου N το πλήθος των χρησιμοποιούμενων χαρακτηριστικών. Παραδείγματα τέτοιων χαρακτηριστικών είδαμε στην παράγραφο 2.1.2. Η διαφορετικότητα της προσέγγισής μας επιτάσσει τη χρήση ενός συνόλου χαρακτηριστικών πολλά από τα οποία διαφέρουν σημαντικά από όσα συναντήσαμε στη βιβλιογραφία.

Πριν προχωρήσουμε στην μελέτη των χαρακτηριστικών που επιλέχθηκαν, υπενθυμίζουμε ότι καθένα από αυτά πρέπει να έχει νόημα για την περιγραφή όχι μόνο των ταινιών, αλλά επίσης των ηθοποιών και των σκηνοθετών.

Εφόσον δεν είναι δυνατή η δημιουργία χαρακτηριστικών που να εκφράζουν την ομοιότητα ερωτήματος-αποτελέσματος όπως εξηγήθηκε στην παράγραφο 5.1, τα χαρακτηριστικά που δημιουργήθηκαν αφορούν το ιστορικό του χρήστη (προγενέστερες βαθμολογίες ταινιών) και τη δομή του rdf γράφου. Τα χαρακτηριστικά που δημιουργήθηκαν είναι τα εξής:

1. Ηθοποιοί

   Καθένα από τα χαρακτηριστικά αυτής της κατηγορίας αντιστοιχεί σε έναν από το σύνολο των ηθοποιών που συμμετέχουν στις 5179 ταινίες του συνόλου των δεδομένων μας. Πρόκειται για χαρακτηριστικά με boolean τιμές (δηλαδή 0 ή 1), οι οποίες ορίζονται ως εξής:

   Για τις ταινίες, ένα χαρακτηριστικό έχει τιμή ίση με 1 εάν πρόκειται για ηθοποιό που συμμετέχει σε αυτή και 0 σε άλλη περίπτωση.

   Για τους ηθοποιούς όλα τα χαρακτηριστικά έχουν τιμή 0 εκτός εκείνο που τους αντιστοιχεί.

2. Σκηνοθέτες

   Καθένα από τα χαρακτηριστικά αυτής της κατηγορίας αντιστοιχεί σε έναν από το σύνολο των σκηνοθετών που συναντώνται στις 5179 ταινίες του συνόλου των δεδομένων μας. Πρόκειται για χαρακτηριστικά με boolean τιμές (δηλαδή 0 ή 1), οι οποίες ορίζονται ως εξής:

   Για τις ταινίες, ένα χαρακτηριστικό έχει τιμή ίση με 1 εάν πρόκειται για το σκηνοθέτη της και 0 σε άλλη περίπτωση.

Για τους σκηνοθέτες όλα τα χαρακτηριστικά έχουν τιμή 0 εκτός από εκείνο που τους αντιστοιχεί.

3. SKOS κατηγορίες για ταινίες

Καθένα από τα χαρακτηριστικά αντιστοιχεί σε μια κατηγορία του λεξιλογίου SKOS από εκείνες που επιλέχθηκαν για τις ταινίες σύμφωνα με το κεφάλαιο 4. Αποκλείστηκαν, όμως, πρώτα κάποιες κατηγορίες που σχετίζονται με ταινίες αλλά δεν κρίθηκαν σημαντικές και επιπλέον, όσες κατηγορίες συγκέντρωναν λιγότερες από 5 ταινίες. Δεδομένου ότι το σύνολο των δεδομένων μας περιλαμβάνει 5179 ταινίες, δεν πρόκειται για αυστηρό κατώφλι, δε χάνουμε δηλαδή σημαντικές κατηγορίες.

Μια ταινία έχει τιμή 1 για κάποιο χαρακτηριστικό αν ανήκει στην κατηγορία που αντιπροσωπεύει. Διαφορετικά έχει τιμή 0.

Για τους ηθοποιούς και τους σκηνοθέτες το χαρακτηριστικό έχει τιμή ίση με το πλήθος των ταινιών στις οποίες συμμετείχαν και οι οποίες ανήκουν στην κατηγορία που αντιστοιχεί το χαρακτηριστικό.

Πρέπει στο σημείο αυτό να κάνουμε την εξής επισήμανση. Γνωρίζουμε πως η γνώση ότι μια ταινία ανήκει σε μια κατηγορία προέρχεται από το αρχείο *** της DBpedia ή μέσα από την ιεραρχία κατηγοριών όπως περιγράφηκε στο κεφάλαιο 4. Θα ήταν εύλογη η αξιοποίηση της διαφορετικής προέλευσης της πληροφορίας για την απόδοση διαφορετικών τιμών σε κάποιο χαρακτηριστικό της κατηγορίας. Για παράδειγμα μια ταινία μπορεί να ανήκει άμεσα στην κατηγορία <Teen Comedy Films> και έμμεσα στην κατηγορία <Comedy Films>. Θεωρήθηκε όμως ότι πρόκειται ουσιαστικά για κατηγορίες του ίδιου επιπέδου και άρα δεν έχει νόημα η προώθηση κάποιας ως πιο σημαντικής ή αντιπροσωπευτικής για την ταινία. Ο χαλαρός τρόπος ορισμού της ιεραρχίας θα δημιουργούσε άλλωστε προβλήματα στον ορισμό κανόνων για απόδοση τιμών με κριτήριο το βάθος στο οποίο συναντάται μια κατηγορία. Παράλληλα, επειδή τα χρησιμοποιούμενα χαρακτηριστικά πρέπει να βρίσκουν επίσης εφαρμογή σε ηθοποιούς και σκηνοθέτες, με την πρακτική αυτή δεν είναι προφανές ποια θα ήταν εδώ η σωστή απόδοση τιμών. Επιπλέον, δεν είναι απαραίτητα σωστή η θεώρηση ότι η πιο ειδική πληροφορία είναι πιο ενημερωτική για τις προτιμήσεις του χρήστη. Παρόλο που οι βραβευμένες με Oscar/Χρυσή Σφαίρα ταινίες είναι υποκατηγορίες των ταινιών που έχουν λάβει βραβείο, δεν είναι σαφές αν για το χρήστη βαρύνει περισσότερο κάποιο συγκεκριμένο βραβείο ή η απόκτηση ενός οποιουδήποτε βραβείου για να θεωρήσει την ταινία πιο αξιόλογη. Τέλος, εξαιτίας του τρόπου με τον οποίο συλλέγεται η γνώση στα άρθρα της Wikipedia, δεν μπορούμε να θεωρήσουμε αυστηρά σωστές τις διαθέσιμες πληροφορίες, δεν ξέρουμε δηλαδή αν μια ταινία ανήκει στην κατηγορία "Comedy" γιατί δεν υπάρχει πιο εξειδικευμένη κατηγορία να της αρμόζει ή επειδή αυτή η πληροφορία δεν

είναι διαθέσιμη. Για όλους τους παραπάνω λόγους υιοθετήθηκε τελικά η boolean λογική στις τιμές των χαρακτηριστικών αυτών.

4. SKOS κατηγορίες για ηθοποιούς και σκηνοθέτες

   Καθένα από τα χαρακτηριστικά αντιστοιχεί σε μια κατηγορία του λεξιλογίου SKOS από εκείνες που επιλέχθηκαν για τους συντελεστές των ταινιών σύμφωνα με το κεφάλαιο 4. Αποκλείστηκαν, όμως, πρώτα κάποιες κατηγορίες που σχετίζονται με συντελεστές ταινιών αλλά δεν κρίθηκαν σημαντικές και επιπλέον, όσες κατηγορίες συγκέντρωναν λιγότερους από 10 συντελεστές. Το μεγαλύτερο κατώφλι συγκριτικά με αυτό που χρησιμοποιήθηκε στις ταινίες, επιλέχθηκε γιατί το ενδιαφέρον που συγκεντρώνουν οι διάφοροι συντελεστές προήλθε έμμεσα μέσω των βαθμολογιών των ταινιών στις οποίες συμμετείχαν και κρίθηκε καταλληλότερη μια αυστηρότερη διαδικασία επιλογής των σημαντικών χαρακτηριστικών.

   Μια ταινία έχει τιμή ίση με το πλήθος των συντελεστών της οι οποίοι ανήκουν στην κατηγορία που αντιστοιχεί το χαρακτηριστικό.

   Για τους ηθοποιούς και τους σκηνοθέτες το χαρακτηριστικό έχει τιμή ίση με 1 αν ανήκουν στην κατηγορία που αντιπροσωπεύει. Διαφορετικά λαμβάνει την τιμή 0.

5. Είδη ταινιών

   Καθένα από τα χαρακτηριστικά αντιπροσωπεύει ένα είδος ταινίας από αυτά του πίνακα 8.

   Μια ταινία έχει τιμή 1 για κάποιο χαρακτηριστικό αν ανήκει στο συγκεκριμένο είδος. Διαφορετικά έχει τιμή 0.

   Για τους ηθοποιούς και τους σκηνοθέτες το χαρακτηριστικό έχει τιμή ίση με το πλήθος των ταινιών στις οποίες συμμετείχαν και οι οποίες ανήκουν στο συγκεκριμένο είδος.

6. Βαθμός του κόμβου στον RDF γράφο

   Χρησιμοποιείται ως μέτρο της δημοφιλίας μιας ταινίας ή ενός συντελεστή της και ισούται με το πλήθος των σχέσεων (τριπλέτες) στις οποίες η εν λόγω οντότητα (ταινία, ηθοποιός ή σκηνοθέτης) συμμετέχει είτε ως υποκείμενο είτε ως αντικείμενο

7. Πλήθος σχέσεων μεταξύ ταινιών και συντελεστών

   Για τις ταινίες η τιμή του χαρακτηριστικού είναι ίση με τον αριθμό των συντελεστών της που είναι ηθοποιοί ή σκηνοθέτες.

   Για τους ηθοποιούς και τους σκηνοθέτες η τιμή του ισούται με τις ταινίες στις οποίες έχουν παίξει ή τις οποίες έχουν σκηνοθετήσει αντίστοιχα. Σε περίπτωση που ένας

ηθοποιός έχει εργαστεί και ως σκηνοθέτης (και αντίστροφα), η τιμή του χαρακτηριστικού είναι το άθροισμα των ταινιών στις οποίες συμμετείχε με καθέναν από τους δύο ρόλους.

8. YAGO κλάσεις για ταινίες

   Καθένα από τα χαρακτηριστικά αντιστοιχεί σε μια κλάση της οντολογίας YAGO όπως περιγράφηκε στο κεφάλαιο 4.

   Μια ταινία έχει τιμή 1 για κάποιο χαρακτηριστικό αν ανήκει στην κλάση που αντιπροσωπεύει. Διαφορετικά λαμβάνει τιμή 0.

   Για τους ηθοποιούς και τους σκηνοθέτες το χαρακτηριστικό έχει τιμή ίση με το πλήθος των ταινιών στις οποίες συμμετείχαν και οι οποίες ανήκουν στην κλάση στην οποία αντιστοιχεί το χαρακτηριστικό.

9. YAGO κλάσεις για ηθοποιούς και σκηνοθέτες

   Καθένα από τα χαρακτηριστικά αντιστοιχεί σε μια κλάση της οντολογίας YAGO όπως περιγράφηκε στο κεφάλαιο 4.

   Μια ταινία έχει τιμή ίση με το πλήθος των συντελεστών της οι οποίοι ανήκουν στην κλάση στην οποία αντιστοιχεί το χαρακτηριστικό.

   Για τους ηθοποιούς και τους σκηνοθέτες το χαρακτηριστικό έχει τιμή ίση με 1 αν ανήκουν στην κλάση που αντιπροσωπεύει. Διαφορετικά λαμβάνει την τιμή 0.

10. Αντιστοίχηση είδους ταινίας με SKOS κατηγορία

    Κάθε χαρακτηριστικό της κατηγορίας αντιπροσωπεύει, όπως και στο 5, ένα είδος ταινίας. Στο κεφάλαιο 3 παρουσιάσθηκε η ανάκτηση της σχετικής με τα είδη πληροφορίας από τα αρχεία του IMDB. Καθώς δεν ήταν δυνατή η αντιστοίχιση όλων των 5179 ταινιών, για πολλές ταινίες δεν υπάρχει πληροφορία για το είδος που ανήκουν. Ωστόσο, πολλές SKOS κατηγορίες αφορούν τα είδη των ταινιών. Είναι προφανές πως δύο ταινίες που ανήκουν στην κατηγορία "Κωμωδία", θα πρέπει να μοιράζονται αυτή την ιδιότητα, ανεξάρτητα από την προέλευση της πληροφορίας. Για το λόγο αυτό, όπου ήταν δυνατό, έγιναν οι αντιστοιχίσεις από τα είδη του IMDB στα είδη των SKOS κατηγοριών. Οι αντιστοιχίσεις που προέκυψαν είναι οι εξής:

| | |
|---|---|
| Crime | Category:Crime_films |
| HOrrOr | Category:Horror_films |
| ROmance | Category:Romance_films |
| Mystery | Category:Mystery_films |
| COmedy | Category:Comedy_films |

| | |
|---|---|
| SpOrt | Category:Sports_films |
| Fantasy | Category:Fantasy_films |
| War | Category:War_films |
| ActiOn | Category:Action_films |
| Musical | Category:Musical_films |
| Thriller | Category:Thriller_films |
| DOcumentary | Category:Documentary_films |
| Western | Category:Western_films |
| Adventure | Category:Adventure_films |
| Film-NOir | Category:Film_noir |
| Sci-Fi | Category:Science_fiction_films |
| Drama | Category:Drama_films |

**Πίνακας 11: Αντιστοίχιση IMDB genre σε SKOS category**

Η κατασκευή των χαρακτηριστικών αυτής της κατηγορίας κατέστη δυνατή χάρη στην αξιοποίηση της SKOS ιεραρχίας όπως παρουσιάστηκε σε προηγούμενο κεφάλαιο.

Οι τιμές για τα χαρακτηριστικά της τελευταίας αυτής κατηγορίας ανατίθενται όπως στην κατηγορία 5 (είδη ταινιών), με μόνη διαφορά ότι μια ταινία θεωρείται τώρα ότι ανήκει σε ένα είδος και στην περίπτωση που ανήκει στην αντίστοιχη κατηγορία SKOS.

ΠΑΡΑΤΗΡΗΣΗ

Ενώ οι ταινίες που μας ενδιαφέρουν περιορίζονται στο σύνολο των 5179 ταινιών στο οποίο καταλήξαμε στο κεφάλαιο 3, ένας ηθοποιός είναι πιθανό να έχει συμμετάσχει και σε ταινίες έξω από το σύνολο αυτό. Στα χαρακτηριστικά που αφορούν το πλήθος των ταινιών στις οποίες συμμετείχε και οι οποίες εμφανίζουν μια συγκεκριμένη ιδιότητα. έχει ληφθεί υπόψη το σύνολο των ταινιών, έτσι ώστε να μην αλλοιώνεται το προφίλ του. Ανεξάρτητα από το αν στο δικό μας σύνολο δεδομένων περιλαμβάνεται ένας μικρός, σχετικά, αριθμός ταινιών, οι χρήστες του Netflix έχουν στη διάθεσή τους στοιχεία για το σύνολο της δουλειάς ενός ηθοποιού. Άλλωστε δεν πρέπει να παραβλεφθεί το γεγονός ότι κατά την αντιστοίχιση των δεδομένων του Netflix στα στοιχεία της DBpedia, η οποία δεν ήταν πλήρης, ένα σημαντικό ποσοστό ταινιών δε διατηρήθηκε. Αντίστοιχα συμπεράσματα ισχύουν και για τους σκηνοθέτες.

Ακόμα, πρέπει να τονιστεί η πιθανή χρονική ανακολουθία στα δεδομένα όπως έχουν διαμορφωθεί σήμερα συγκριτικά με την περίοδο που εξετάζεται στα αρχεία του Netflix. Οποιαδήποτε πληροφορία αφορά ταινίες πέρα από την τελευταία ημερομηνία που δόθηκαν βαθμολογίες από κάποιο χρήστη, δε θα έπρεπε να θεωρείται διαθέσιμη. Λαμβάνεται, δηλαδή, υπόψη, η μετέπειτα πορεία κάποιου συντελεστή, η οποία φυσικά δε θα μπορούσε να έχει επηρεάσει μια προηγούμενη βαθμολόγηση. Είναι όμως ασαφές ποια τμήματα του RDF γράφου πρέπει να αφαιρεθούν και πόσο έγκυρα θα είναι τα εναπομείναντα τμήματα σε

περίπτωση μιας τέτοιας ενέργειας. Το σημαντικότερο ζήτημα που προκύπτει είναι η ένταξη των ηθοποιών και των σκηνοθετών σε κατηγορίες. Για παράδειγμα, για έναν ηθοποιό που ανήκει στην κατηγορία των βραβευμένων με Oscar, δεν είναι εφικτό από τα αρχεία της DBpedia να διαπιστωθεί πότε έλαβε το εν λόγω βραβείο, γιατί η αντίστοιχη πληροφορία δεν είναι διαθέσιμη. Επειδή δεν καθίσταται δυνατός ο σωστός αποκλεισμός πληροφοριών, διατηρούμε τα δεδομένα ως έχουν.

## *5.4 Χαρακτηριστικά εκπαίδευσης που μελετήθηκαν και απορρίφθηκαν*

1. Ποσοστά αντί για απόλυτες τιμές στις ομάδες χαρακτηριστικών 3,4,5,8,9,10

   Εκτός από τον αριθμό των δραματικών ταινιών στις οποίες έχει συμμετάσχει ένας ηθοποιός, θα είχε ίσως ενδιαφέρον η χρήση και του ποσοστού των ταινιών αυτών επί του συνόλου των ταινιών στις οποίες έχει παίξει. Αντίστοιχα και για σκηνοθέτες.

   Αφενός, δεν είναι προφανές τι τιμή πρέπει να πάρει το χαρακτηριστικό όταν ένας ηθοποιός είναι και σκηνοθέτης. Αφετέρου, πρόκειται για πληροφορία που είναι έμμεσα διαθέσιμη από το πλήθος των δραματικών ταινιών και το συνολικό αριθμό των ταινιών στις οποίες συμμετείχε. Το βασικότερο πρόβλημα είναι η αδυναμία ανάθεσης κατάλληλης τιμής στο χαρακτηριστικό αυτό για μια ταινία.

2. Budget

   Ο προϋπολογισμός των ταινιών που έχει βαθμολογήσει ο χρήστης θα μπορούσε να χρησιμοποιηθεί για την ανίχνευση προτιμήσεων σε ανεξάρτητες παραγωγές και εναλλακτικό κινηματογράφο (κατεξοχήν χαμηλού προϋπολογισμού) ή σε ταινίες των κυρίαρχων τάσεων (mainstream) και ταινίες με πολλά ειδικά εφέ (κατεξοχήν υψηλού προϋπολογισμού). Τα στοιχεία όμως που παρέχονται από τη DBpedia κρίθηκαν ιδιαίτερα αναξιόπιστα και η χρήση τους απορρίφθηκε. Η χρήση στα infoboxes της Wikipedia άλλοτε ολόκληρων τιμών και άλλοτε τιμών εκφρασμένων σε εκατομμύρια έχει ως αποτέλεσμα ανακολουθίες μεταξύ των δεδομένων.

3. Gross

   Οι εισπράξεις μιας ταινίας επίσης θα μπορούσαν να χρησιμοποιηθούν για να διαπιστωθεί αν ο χρήστης προτιμά περισσότερο ή λιγότερο εμπορικές ταινίες. Όπως και στην περίπτωση του προϋπολογισμού, μετά από στατιστικές αναλύσεις και δειγματοληπτική εξέταση κάποιων ταινιών, η χρήση αυτού του χαρακτηριστικού απορρίφθηκε εξαιτίας της αναξιοπιστίας των δεδομένων.

4. Δεκαετία στην οποία ανήκει μια ταινία

   Η πληροφορία αυτή είναι διαθέσιμη από τις κατηγορίες χαρακτηριστικών 3 και 8 και δεν κρίθηκε σκόπιμη η δημιουργία ξεχωριστών χαρακτηριστικών.

## *5.5 Συνδυασμός ομάδων χαρακτηριστικών*

Οι δέκα ομάδες χαρακτηριστικών που περιγράφηκαν παραπάνω δημιουργήθηκαν με σκοπό την πλήρη αξιοποίηση της διαθέσιμης πληροφορίας για το σύνολο των ταινιών. Καθώς η προσέγγιση που ακολουθείται εδώ δε συναντάται στη βιβλιογραφία, κατεβλήθη προσπάθεια τα χαρακτηριστικά που θα χρησιμοποιηθούν να παρουσιάζουν μεγάλη ποικιλομορφία ώστε να αυξηθεί η πιθανότητα με κατάλληλο συνδυασμό τους να εκπαιδευτεί σωστά το SVM. Το πλήθος τους καθώς και η επανάληψη της πληροφορίας σε κάποια από αυτά (π.χ. κατηγορίες 3 και 8), επιτρέπει το συνδυασμό κατηγοριών χαρακτηριστικών ώστε να μελετηθεί ποια ομάδα χαρακτηριστικών είναι η πλέον κατάλληλη για την εκπαίδευση του SVM.

# 6

# *Μηχανή Αναζήτησης*

Η εξατομίκευση των αποτελεσμάτων που παρουσιάζονται στο χρήστη ως απάντηση σε ένα ερώτημά του, προϋποθέτει ασφαλώς την ανάκτηση αυτών των αποτελεσμάτων από τη βάση δεδομένων στην οποία πραγματοποιείται η αναζήτηση. Η διαδικασία της αναταξινόμησης των αποτελεσμάτων ώστε η σειρά τους να ικανοποιεί περισσότερο τις ανάγκες του χρήστη έχει νόημα μόνο αν σε πρώτη φάση η μηχανή αναζήτησης έχει ανακτήσει το σύνολο των σχετικών αποτελεσμάτων και, εφόσον είναι δυνατό, μόνο αυτά. Δε θα είχε νόημα να αλλάζουμε τη σειρά εμφάνισης των αποτελεσμάτων όταν κάποια από αυτά δε θα έπρεπε εξαρχής να έχουν συμπεριληφθεί. Αντίστοιχα, αν από τα επιστρεφόμενα αποτελέσματα λείπουν σημαντικά δεδομένα, η εξατομίκευση δε θα μπορέσει να αναπληρώσει το κενό και ο χρήστης ίσως δε βρει αυτό που τον ενδιαφέρει. Η μηχανή αναζήτησης πρέπει να έχει ταυτόχρονα υψηλή ανάκληση αλλά και ακρίβεια.

Η αναζήτηση με λέξεις-κλειδιά σε σημασιολογικά δεδομένα είναι ένα σχετικά νέο πεδίο μελέτης. Ο τρόπος με τον οποίο γίνεται η αναζήτηση ώστε να είναι αποδοτική, καθώς και η μορφή των επιστρεφόμενων αποτελεσμάτων, δεν είναι σαφώς καθορισμένα και εξαρτώνται από τις ανάγκες που έρχεται να καλύψει κάθε εφαρμογή.

Για τις ανάγκες των δικών μας πειραμάτων θα χρησιμοποιήσουμε μια στοιχειώδη μηχανή αναζήτησης σε RDF που έχει δημιουργηθεί στο Ινστιτούτου Πληροφοριακών Συστημάτων (ΙΠΣΥ). Το παρόν κεφάλαιο είναι αφιερωμένο στην παρουσίαση αυτής της μηχανής αναζήτησης και των επεκτάσεων που πραγματοποιήσαμε.

## *6.1 Τρόπος λειτουργίας εφαρμογής*

Η εφαρμογή παρέχει τη δυνατότητα αναζήτησης με λέξεις-κλειδιά σε αρχεία με N-TRIPLE format. Η εφαρμογή παίρνει ως είσοδο:

- Τις φράσεις-κλειδιά που θέλει ο χρήστης να αναζητήσει
- Το πλήθος (INITIALRESULTSIZE) των αρχικών αποτελεσμάτων που θα συγκεντρώσει η μηχανή για καθεμία από τις φράσεις-κλειδιά (προαιρετικά).
- Το πλήθος (TOPRESSIZE) των αποτελεσμάτων που θα κρατήσει η μηχανή για καθεμιά από τις φράσεις-κλειδιά (προαιρετικά).
- Το μέγιστο μήκος μονοπατιού (PLENGTH) που μπορεί να συνδέει 2 φράσεις-κλειδιά (προαιρετικά).

Αν δεν δοθούν οι προαιρετικές παράμετροι, τα αντίστοιχα μεγέθη παίρνουν κάποιες προκαθορισμένες από τη μηχανή αναζήτησης τιμές.

Οι διαφορετικές φράσεις-κλειδιά δίνονται διαχωρισμένες με "," από το χρήστη, οπότε η μηχανή αναζήτησης δεν επιβαρύνεται με την προσπάθεια ομαδοποίησης των λέξεων που θα αύξανε το πλήθος των απαραίτητων αναζητήσεων. Για παράδειγμα, το ερώτημα "obama white house" θα δημιουργούσε 7 διαφορετικούς συνδυασμούς λέξεων για τους οποίους η μηχανή θα εκτελούσε ξεχωριστή αναζήτηση, ενώ το ερώτημα "obama, white house" απαιτεί την εκτέλεση 2 μόνο αναζητήσεων.

Στο πρώτο στάδιο, με χρήση μεθόδων που παρέχει το Apache Lucene, για κάθε φράση-κλειδί ανακτώνται από μια βάση δεδομένων όλα τα αποτελέσματα που βρίσκονται να σχετίζονται με αυτή. Ως αποτελέσματα νοούνται μόνο κόμβοι του γράφου δεδομένων. Ένας κόμβος μπορεί να ανακτηθεί αν το όνομά του (label) ταιριάζει με τη φράση ή αν το όνομα κάποιου συσχετισμένου με αυτόν κόμβου (π.χ. το label της κλάσης στην οποία ανήκει) ταιριάζει με τη φράση. Στο στάδιο αυτό δε μας επηρεάζει το πόσο σχετικά είναι τα αποτελέσματα. Κρατάμε κάθε σχετικό αποτέλεσμα μέχρι να συμπληρωθούν INITIALRESULTSIZE αποτελέσματα ή να μην υπάρχει κανένα άλλο σχετικό αποτέλεσμα. Αν μάλιστα ένας κόμβος ανακτήθηκε παραπάνω από μια φορά, τότε προστίθενται στα αποτελέσματα ισάριθμες εγγραφές URI κόμβου-λόγος που ανακτήθηκε-κειμενική ομοιότητα(score).

Η λίστα αποτελεσμάτων που έχει δημιουργηθεί για κάθε φράση-κλειδί, αναδιατάσσεται με βάση το σκορ κάθε αποτελέσματος, γίνεται απαλοιφή διπλότυπων (ίδιος κόμβος, άλλο score) και στη συνέχεια κρατάμε τα TOPRESSIZE αποτελέσματα με το μεγαλύτερο score.

Τέλος, επιλέγονται ανά δύο οι φράσεις-κλειδιά και η μηχανή επιχειρεί να συνδέσει καθένα από τα (το πολύ) TOPRESSIZE αποτελέσματα της πρώτης φράσης με καθένα από τα (το

πολύ) TOPRESSIZE της δεύτερης φράσης με μονοπάτια μήκους το πολύ PLENGTH αναζητώντας σε προκαθορισμένα RDF αρχεία. Για το σκοπό αυτό χρησιμοποιείται το Jena API.Τα μονοπάτια αυτά είναι η έξοδος της εφαρμογής.

## *6.2 Επέκταση λειτουργίας της μηχανής αναζήτησης*

Πριν αναφερθούμε στις τροποποιήσεις και επεκτάσεις που πραγματοποιήσαμε στη λειτουργία της μηχανής αναζήτησης, πρέπει να καθοριστεί το σύνολο των δεδομένων που θα αποτελέσουν τη βάση της αναζήτησής μας, έτσι ώστε να αποκτήσουμε μια εικόνα του όγκου των δεδομένων που θα πρέπει να χειρίζεται η εφαρμογή.

Για την αρχική αναζήτηση αποτελεσμάτων με τη μορφή απλών κόμβων που σχετίζονται με τις λέξεις-κλειδιά χρησιμοποιούνται οι πληροφορίες από τα εξής αρχεία της DBpedia :

- Ontology Infobox Types
- Ontology Infobox Properties (επιλεγμένες εγγραφές)
- Titles
- Articles Categories
- Categories (Labels)

Χρειάστηκε να κάνουμε αλλαγές στους τίτλους (labels) κάποιων εγγραφών στις περιπτώσεις που σε αυτούς περιλαμβάνονταν στοιχεία για την κλάση στην οποία ανήκει η εγγραφή σύμφωνα με την οντολογία της dbpedia. Η αλλαγή αυτή εξασφαλίζει ότι θα δίνεται σε όλες τις οντότητες μιας κλάσης η ίδια βαθμολογία (score) από τη μηχανή αναζήτησης αν ως λέξη κλειδί χρησιμοποιηθεί το όνομα της κλάσης τους . Η αναζήτηση μονοπατιών που συνδέουν μεταξύ τους δύο κόμβους γίνεται στις τριπλέτες των αρχείων:

- Articles Categories
- Ontology Infobox Properties

αφού αφαιρεθούν οι σχέσεις που έχουν literals ως αντικείμενα.

Το μεγάλο μέγεθος της βάσης δεδομένων στην οποία πραγματοποιείται η αναζήτηση σε συνδυασμό με την ιδιαίτερα ευέλικτη ανάκτηση πιθανών αποτελεσμάτων δημιουργούν προβλήματα στην απόδοση της μηχανής αναζήτησης.

### *6.2.1 Επιλογή καλύτερων υποψήφιων αποτελεσμάτων στις αρχικές λίστες κάθε όρου αναζήτησης*

Ένας προφανής τρόπος ανάκτησης αποτελεσμάτων θα ήταν αρχικά να αντιστοιχηθεί κάθε φράση-κλειδί σε έναν κόμβο του γράφου και στη συνέχεια να βρεθούν τα μονοπάτια που

συνδέουν αυτούς τους κόμβους. Ωστόσο η μηχανή αναζήτησης που διαθέτουμε επιχειρεί να ανακτήσει όλους τους κόμβους που σχετίζονται με μια φράση-κλειδί έτσι ώστε να αυξηθεί η πιθανότητα να βρεθεί το αποτέλεσμα που επιθυμεί ο χρήστης. Ως ένα παράδειγμα των δυνατοτήτων αυτής της πρακτικής, εξετάζουμε τη φράση-κλειδί "Woody Allen". Ενώ η προφανής επιλογή είναι η ανάκτηση του κόμβου το URI του οποίου αντιστοιχεί στην εγγραφή της DBpedia για το Woody Allen, στη δική μας μηχανή αναζήτησης, εκτός από αυτόν τον κόμβο, θα εμφανιστούν στα αποτελέσματα και ταινίες του σκηνοθέτη, σε χαμηλότερη φυσικά θέση. Η επιλογή αυτή, σε πολύ αφηρημένες λέξεις-κλειδιά, μπορεί να βελτιώσει σημαντικά την ποσότητα και την ποιότητα των επιστρεφόμενων αποτελεσμάτων. Ένας απλός χρήστης θα περίμενε, για παράδειγμα, η λέξη "film" να επιστρέφει ταινίες. Αν όμως η μηχανή πρέπει να επιστρέψει αυστηρά ένα αποτέλεσμα από τη DBpedia, αυτό θα είναι ο κόμβος με URI <http://dbpedia.org/ontology/Film>, πιθανότατα όχι αυτό που αναζητούσε ο χρήστης. Υπενθυμίζουμε ότι σκοπός μας είναι η εφαρμογή μας να μπορεί να χρησιμοποιηθεί και από μη εξοικειωμένους με το RDF μοντέλο χρήστες.

Με τον τρόπο που περιγράφηκε, παρέχεται μεγαλύτερη ευελιξία στην αναζήτηση, αφού ανακτώνται όλες οι εγγραφές που σχετίζονται με οποιαδήποτε από τις φράσεις-κλειδιά. Τα τελικά αποτελέσματα που θα επιστραφούν στο χρήστη δεν αποτελούνται όμως από αυτές τις ταξινομημένες λίστες εγγραφών. Τα τελικά αποτελέσματα έχουν τη μορφή γράφων που συνδέουν μεταξύ τους τις λέξεις-κλειδιά, τους κόμβους δηλαδή που έχουν ανακτηθεί για καθεμία. Έστω ότι έχουμε 2 λέξεις-κλειδιά και ορίζουμε το TOPRESSIZE ίσο με 10 και το PLENGTH ίσο με 2. Στην απόλυτα ρεαλιστική περίπτωση που τα διακριτά αποτελέσματα για κάθε λέξη είναι τουλάχιστον ίσα με TOPRESSIZE, η μηχανή αναζήτησης πρέπει να εξετάσει 100 συνδυασμούς κόμβων και να αναζητήσει όλα τα μονοπάτια μήκους το πολύ 2 που να τους ενώνουν. Διαφαίνεται ήδη το μεγάλο υπολογιστικό φορτίο που συνεπάγεται η παρεχόμενη ευελιξία. Καθώς μάλιστα ο όγκος των δεδομένων που φιλοξενεί η DBpedia είναι πολύ μεγάλος, αποδεικνύεται ότι τα μεγέθη των αποτελεσμάτων που πρέπει να επιστραφούν και να αξιολογηθούν είναι πολύ μεγαλύτερα.

Για να γίνει αυτό κατανοητό, δίνουμε ένα παράδειγμα ερωτήματος, έστω "Woody Allen, Scarlett Johansson,Film", το οποίο μπορεί να τεθεί από ένα χρήστη που αναζητά ταινίες που να σχετίζονται με τους δύο αυτούς ηθοποιούς και ας θεωρήσουμε ότι έχουμε τη δυνατότητα να ανακτήσουμε απεριόριστο αριθμό κόμβων για κάθε φράση-κλειδί. Για τους δύο ηθοποιούς θα επιστραφούν τα URI που ανήκουν σε αυτούς, σε ταινίες τους και ίσως κάποια λίγα ακόμα αποτελέσματα που σχετίζονται με μη αναμενόμενους τρόπους με αυτούς. Είναι μάλλον προφανές ότι σε αυτό το πλαίσιο η ανάκτηση ταινιών από τα ονόματα των ηθοποιών είναι περιττή. Η παρουσία της λέξης "Film" υποδηλώνει ότι αυτός ο όρος αναζήτησης αναμένεται να επιστρέψει ταινίες. Επειδή αυτό δε λαμβάνεται υπόψη, η μηχανή αναζήτησης θα φέρει όλα

τα σχετικά αποτελέσματα για κάθε όρο. Αυτό πιθανώς συνεπάγεται ότι η ίδια ταινία μπορεί να έρθει ως αποτέλεσμα και για τα δύο ονόματα και, ακολούθως για το "Film".

Το πραγματικό πρόβλημα όμως είναι άλλο. Για τη λέξη "Film", η ασάφειά της και το γεγονός ότι αποτελεί κλάση οντοτήτων, θα υποχρεώσει τη μηχανή αναζήτησης να φέρει όλα τα αντικείμενα της κλάσης, περίπου δηλαδή 60000 URI ταινιών. Τα πραγματικά μεγέθη, εξαιτίας της ανάκτησης κόμβων και μέσω των SKOS κατηγοριών στις οποίες ανήκουν, είναι μεγαλύτερα. Ήδη όμως το πρόβλημα είναι εμφανές. Είναι πρακτικά αδύνατο να εξετάσουμε όλα αυτά τα αποτελέσματα για το σχηματισμό μονοπατιών που να τα συνδέουν με τους άλλους κόμβους.

Αναγνωρίζουμε ότι ο αυθαίρετος αποκλεισμός αποτελεσμάτων δεν μπορεί να είναι λύση. Για να αντιμετωπίσουμε το πρόβλημα που περιγράψαμε, είναι βέβαια επιτακτική η ανάγκη μείωσης των επιστρεφόμενων αποτελεσμάτων, αφού όμως εντοπίσουμε ποια αποτελέσματα δε θα μας φανούν χρήσιμα. Για το σκοπό αυτό, προχωράμε στις εξής δύο ενέργειες:

α) Ορισμός κατωφλίου στο σκορ των αποτελεσμάτων που κρατάμε

Ξέρουμε ότι κάθε αποτέλεσμα συνοδεύεται από ένα σκορ που φανερώνει την ομοιότητά του με τους όρους αναζήτησης (κειμενική ομοιότητα), άρα τα αποτελέσματα που βρίσκονται που βρίσκονται ψηλά στη λίστα είναι και τα πιο σχετικά. Θα θέσουμε ένα όριο ομοιότητας κάτω από το οποίο το αποτέλεσμα δε θα θεωρείται ικανοποιητικά σχετικό με τον όρο αναζήτησης.

Ο τρόπος με τον οποίο το Lucene αναθέτει τα σκορ στα διάφορα αποτελέσματα κάνει χρήση τεχνικών ομοιότητας κειμένου. Τα ίδια τα σκορ δε φέρουν καμία πληροφορία, παρά μόνο ποια αποτελέσματα είναι καλύτερα από τα άλλα. Έχουν νόημα δηλαδή μόνο για τη σύγκριση τιμών και όχι ως απόλυτες τιμές. Άρα δεν μπορούμε να χρησιμοποιήσουμε κάποιο συγκεκριμένο σκορ ως κατώφλι. Αντί για αυτό, χρησιμοποιούμε ως κατώφλι ένα προκαθορισμένο ποσοστό του μεγαλύτερου σκορ που έχει εμφανιστεί στα αποτελέσματα για το συγκεκριμένο όρο αναζήτησης. Το όριο αυτό ορίζεται μεταξύ 50% και 83% του μεγαλύτερου σκορ, αποκλείοντας με αυτό τον τρόπο αποτελέσματα που παρουσιάζουν σημαντικά μικρότερο βαθμό συσχέτισης με το ερώτημα σε σύγκριση με άλλα αποτελέσματα που έχουν ανακτηθεί. Το ακριβές κατώφλι επιλέγεται σε συνάρτηση με το ερώτημα που έχει τεθεί κάθε φορά από το χρήστη.

Έτσι, στο παράδειγμά μας, με κατώφλι 50% , για τη φράση Woody Allen θα κρατήσουμε το URI που του αντιστοιχεί, URI ταινιών του, αλλά όχι το http://dbpedia.org/resource /A._A._Allen.

β) Χρήση του αρχείου Extended Abstracts στην αναζήτηση

Ενώ για τη φράση Woody Allen μπορούμε να ισχυριστούμε ότι έχει νόημα να κρατήσουμε ένα σχετικά μικρό πλήθος αποτελεσμάτων ή ακόμα και ένα μοναδικό αποτέλεσμα, δεν ισχύει το ίδιο για τα αποτελέσματα που αφορούν γενικότερου περιεχομένου λέξεις-κλειδιά. Στην περίπτωση ανάκτησης αποτελεσμάτων μέσω της κλάσης στην οποία ανήκουν (ή κάποιας SKOS κατηγορίας που τα περιέχει), όλα τα αποτελέσματα έχουν το ίδιο σκορ. Στο παράδειγμά μας, καμιά ταινία δε θα υπερτερεί σε σχέση με τις άλλες για τη λέξη-κλειδί "Film". Άρα ούτε θέλουμε ούτε και μπορούμε να αποκλείσουμε κάποια αποτελέσματα με χρήση κατωφλίου.

Στο σημείο αυτό, ο μόνος τρόπος να διαχωρίσουμε τις ταινίες σε περισσότερο και λιγότερο σχετικές, είναι να αξιοποιήσουμε τα συμφραζόμενα της λέξης "Film", δηλαδή τις άλλες δύο φράσεις του ερωτήματος. Αφού τελικός στόχος μας είναι να βρούμε μονοπάτια που να συνδέουν τα αποτελέσματα για τους διάφορους όρους του ερωτήματος, θα πρέπει να προωθήσουμε τα αποτελέσματα εκείνα τα οποία έχουμε λόγο να πιστεύουμε πως πράγματι συνδέονται με τα άλλα. Για να το πετύχουμε αυτό, εισάγουμε ένα ακόμα επίπεδο αναζήτησης.

Συγκεκριμένα, χρησιμοποιούμε το αρχείο της DBpedia που περιέχει για κάθε εγγραφή της αποσπάσματα (abstracts) από το αντίστοιχο άρθρο της Wikipedia. Αν στο abstract ενός αποτελέσματος που αντιστοιχεί σε κάποια λέξη-κλειδί περιέχονται μια ή περισσότερες από τις άλλες λέξεις-κλειδιά αυξάνονται οι πιθανότητες να πρόκειται για ένα σχετικό με το ερώτημα του χρήστη αποτέλεσμα, οπότε το προωθούμε ψηλότερα στη λίστα αυξάνοντας το σκορ του. Έτσι, στο παράδειγμά μας, ταινίες στις οποίες παίζει η Scarlett Johansson ή και έχει σκηνοθετήσει ο Woody Allen, θα αποκτήσουν μεγαλύτερο σκορ.

Στο σημείο αυτό οφείλουμε να τονίσουμε ότι η λύση αυτή δεν είναι απόλυτα σωστή. Η καταλληλότητά της εξαρτάται και από το λόγο για τον οποίο ο χρήστης θέτει ένα ερώτημα. Αν ο σκοπός είναι να ανακαλύψει νέες, μη αναμενόμενες ίσως πληροφορίες, τότε αυτή η προσέγγιση πιθανώς θα τον απογοητεύσει. Κρίθηκε όμως πιο σημαντικό να εξασφαλίσουμε την επιστροφή αποτελεσμάτων εκεί όπου αυτά υπάρχουν από το να χάσουμε, ίσως, κάποια πιο πολύπλοκα μονοπάτια. Είναι ένας συμβιβασμός που πρέπει να κάνουμε έτσι ώστε στην πλειοψηφία των περιπτώσεων η μηχανή αναζήτησης να συμπεριφέρεται με τον αναμενόμενο τρόπο.

Παράλληλα, ο χρόνος απόκρισης της μηχανής αναζήτησης αυξάνεται σημαντικά εάν το μήκος του μονοπατιού που μας ενδιαφέρει ξεπεράσει τις 3 ακμές. Συνεπώς, για σχετικά μικρό μήκος μονοπατιού όπως αυτά που μας απασχολούν στην παρούσα εργασία, είναι ούτως ή άλλως δύσκολο να εξερευνήσουμε πολύπλοκα και μη αναμενόμενα μονοπάτια, γεγονός που κάνει την προσέγγισή μας με τα abstracts ακόμα πιο φυσιολογική. Ο περιορισμός στο μήκος

του μονοπατιού επιβάλλεται από το Jena API που για αναζήτηση σε μεγάλα μονοπάτια απαιτεί μη αποδεκτά πολύ χρόνο.

Τέλος, πρέπει να αναφέρουμε τις εξής πιθανές παραλλαγές της αναζήτησης στα abstracts:

1. αντιμετώπιση της φράσης-κλειδί σαν ενιαία φράση ή διαχωρισμός της σε επιμέρους λέξεις

Στο παράδειγμά μας η φράση "Scarlett Johansson" είναι ένα όνομα και άρα έχει νόημα να την αναζητάμε ολόκληρη στο κείμενο, χωρίς να παρεμβάλλονται δηλαδή άλλες λέξεις ανάμεσα στο όνομα και στο επώνυμο της ηθοποιού. Για μια άλλη φράση κλειδί όμως, όπως για παράδειγμα "drama film" η επιβολή τόσο αυστηρών περιορισμών θα μπορούσε να οδηγήσει σε λανθασμένο αποκλεισμό πολλών εγγραφών. Για παράδειγμα, η ύπαρξη της φράσης "is a drama mystery film" δε θα θεωρούνταν σχετική με τον όρο της αναζήτησης. Για το λόγο αυτό ορίζουμε δύο τρόπους αναζήτησης στα abstracts, καθένας από τους οποίους επιλέγεται εκεί που η χρήση του θεωρείται πιο κατάλληλη. Ένας τρόπος να επιτύχουμε το ίδιο αποτέλεσμα χωρίς να μεταβάλλουμε εμείς αυτή την παράμετρο, θα ήταν να αναζητάμε αρχικά τη φράση κλειδί στους τίτλους των εγγραφών του dataset και εφόσον διαπιστωθεί ότι πρόκειται για τίτλο (π.χ. όνομα) να ακολουθείται η αυστηρότερη αναζήτηση, διαφορετικά να πραγματοποιείται ο δεύτερος, πιο ευέλικτος τρόπος αναζήτησης. Καθώς όμως σε αυτή τη φάση η μηχανή αναζήτησης χρησιμοποιείται αποκλειστικά για τα δικά μας πειράματα, δεν κρίθηκε σκόπιμο να γίνει μια τέτοια επέκταση.

Αντίστοιχα, ορίστηκαν δύο τρόποι χειρισμού των αποτελεσμάτων της αναζήτησης στα abstracts σε συνδυασμό με τα αρχικά αποτελέσματα. Στην πιο αυστηρή εκδοχή, αποκλείουμε εντελώς τα αποτελέσματα που δεν περιέχουν καμία από τις άλλες φράσεις-κλειδιά στα abstracts τους. Στην άλλη, κρατούνται και αυτά, με χαμηλότερο φυσικά σκορ. Και εδώ η καταλληλότερη μέθοδος εξαρτάται από το συγκεκριμένο ερώτημα και το πλήθος των αποτελεσμάτων που έχουν αρχικά επιστραφεί μέσω του lucene.

### *6.2.2 Επιβολή περιορισμών στο σχηματισμό μονοπατιών*

Η ανάγκη επιβολής περιορισμών στα σχηματιζόμενα μονοπάτια εισάγεται από την πληθώρα των διαφορετικών σχέσεων (predicates) που εμφανίζονται στα αρχεία της DBpedia. Δύο κόμβοι μπορεί να συνδέονται μέσω πολλών μονοπατιών, κάποια από τα οποία δεν έχει νόημα να επιστραφούν ως αποτελέσματα. Βασικότερη αιτία για την ύπαρξη τέτοιων μονοπατιών είναι η επιλογή μας να αναζητήσουμε μονοπάτια και μέσω των κατηγοριών SKOS που βασίζονται σε αντίστοιχες κατηγορίες της Wikipedia.

Όπως ξέρουμε, κάθε οντότητα του συνόλου δεδομένων της DBpedia αντιστοιχεί σε ένα άρθρο της Wikipedia. Τα άρθρα της Wikipedia ομαδοποιούνται με βάση ένα πολύπλοκο σύστημα κατηγοριών. Κάθε άρθρο ανήκει συνήθως σε περισσότερες από μια κατηγορίες. Θα αναφερθούμε σε επόμενο κεφάλαιο στην ιεραρχία κατηγοριών της Wikipedia. Εδώ μας

ενδιαφέρει το γεγονός ότι η πληροφορία αυτή είναι διαθέσιμη και για την αντίστοιχη εγγραφή της DBpedia μέσω της ιδιότητας subject. Στις κατηγορίες αυτές βρίσκουμε και πολλές που μας ενδιαφέρουν και έχει νόημα να εμφανίζονται στο γράφο-αποτέλεσμα, όπως "Supernatural horror films" ή " Best Actor Academy Award Winners ". Αντίθετα, κατηγορίες όπως "Living People" μάλλον δεν προσφέρουν σημαντικές πληροφορίες και δεν έχει νόημα να συμπεριληφθούν στην αναζήτηση. Επειδή, όπως είπαμε, η εξατομίκευση αφορά μόνο αποτελέσματα σχετικά με το χώρο του κινηματογράφου, είναι επιθυμητό να περιορίσουμε τα πιθανά αποτελέσματα και με αυτό το κριτήριο. Σκοπός μας δεν είναι να δημιουργήσουμε μια γενικού σκοπού μηχανή αναζήτησης, αλλά μια εφαρμογή που αναζητά κινηματογραφικές πληροφορίες και τις παρουσιάζει στο χρήστη ταξινομημένες σύμφωνα με τις προτιμήσεις του.

Για το σκοπό αυτό χρησιμοποιούμε τα εξής χαρακτηριστικά για να περιγράψουμε το είδος των ακμών που περιλαμβάνει ένα μονοπάτι:

1. πλήθος εμφανίσεων της ιδιότητας
   http://dbpedia.org/ontology/starring
2. πλήθος εμφανίσεων της ιδιότητας
   http://dbpedia.org/ontology/director
3. πλήθος εμφανίσεων της ιδιότητας
   http://dbpedia.org/ontology/writer
4. πλήθος εμφανίσεων μιας εκ των τριών προηγούμενων ιδιοτήτων
5. πλήθος εμφανίσεων μιας ιδιότητας από το σύνολο των σχετικών με ταινίες ιδιοτήτων που αποτελείται από τις:
   http://dbpedia.org/ontology/starring
   http://dbpedia.org/ontology/director
   http://dbpedia.org/ontology/writer
   http://dbpedia.org/ontology/cinematography
   http://dbpedia.org/ontology/narrator
   http://dbpedia.org/ontology/producer
   http://dbpedia.org/ontology/distributor
   http://dbpedia.org/ontology/editing
   http://dbpedia.org/ontology/editor
   http://dbpedia.org/ontology/musicComposer
   http://dbpedia.org/ontology/basedOn
   http://dbpedia.org/ontology/knownFor
   http://dbpedia.org/ontology/influencedBy
   http://dbpedia.org/ontology/award
   http://dbpedia.org/ontology/influenced
   http://dbpedia.org/ontology/creator
   http://dbpedia.org/ontology/notableWork
   http://dbpedia.org/ontology/previousWork

6. πλήθος εμφανίσεων της ιδιότητας
   http://purl.org/dc/terms/subject
7. πλήθος ιδιοτήτων που δεν ανήκουν σε καμία από τις παραπάνω κατηγορίες

Σε περίπτωση που κάποιο μονοπάτι έχει μηδενική τιμή στα χαρακτηριστικά 5 και 6, τότε δε θεωρείται αποδεκτό αποτέλεσμα.

Επιπλέον, παρέχεται η δυνατότητα εφαρμογής αυστηρότερων κριτηρίων σχετικά με το αν θα ληφθεί υπόψη κάποια ακμή " http://purl.org/dc/terms/subject" στην αύξηση της τιμής του 6. Συγκεκριμένα, οι ακμές αυτές λαμβάνονται υπόψη μόνο αν οι κόμβοι στους οποίους καταλήγουν αντιστοιχούν σε κατηγορίες SKOS οι οποίες πληρούν ορισμένα κριτήρια που αφορούν το βαθμό σχετικότητας μιας κατηγορίας με το χώρο του κινηματογράφου. Για τους σκοπούς της παρούσας διπλωματικής, είναι πάντα ενεργοποιημένη η εφαρμογή του αυστηρότερου αυτού αποκλεισμού μονοπατιών.

### *6.2.3 Συνδυασμός μονοπατιών για το σχηματισμό αποτελεσμάτων σε μορφή γράφων*

Η μηχανή αναζήτησης στην αρχική της μορφή επιστρέφει ως αποτελέσματα μονοπάτια μεταξύ δύο κόμβων. Σε περίπτωση που ο χρήστης δώσει σαν είσοδο περισσότερες από μια λέξεις-κλειδιά, η μηχανή θα τις συνδυάσει ανά δύο και θα προσπαθήσει να συνδέσει κάθε αποτέλεσμα της μιας με κάθε αποτέλεσμα της άλλης. Ζητούμενο του χρήστη όμως είναι το συνολικό αποτέλεσμα που θα πάρει ως απάντηση στο ερώτημά του να σχετίζεται με όλες τις λέξεις κλειδιά, εάν αυτό είναι δυνατό. Ο τυχαίος συνδυασμός μονοπατιών δεν αποτελεί λύση, αφού θέλουμε τα μονοπάτια να έχουν όσο γίνεται μεγαλύτερη συνοχή. Θα θέλαμε να έχουμε ένα μέτρο της ομοιότητας μεταξύ δύο μονοπατιών που να μας βοηθά στο σχηματισμό συνεκτικών, νοηματικά, γράφων. Ο καθορισμός μιας τέτοιας ομοιότητας αφορά και πάλι τις ανάγκες της δικής μας εφαρμογής. Κάθε μονοπάτι αντιστοιχεί σε ένα 7-διάστατο διάνυσμα, οι διαστάσεις του οποίου αντιστοιχούν στα 7 χαρακτηριστικά που ορίσαμε στην προηγούμενη ενότητα. Για το συνδυασμό μονοπατιών, χρησιμοποιούμε cosine similarity μεταξύ των διανυσμάτων που τους αντιστοιχούν.

## *6.3 Παραδείγματα ερωτημάτων και αποτελεσμάτων*

Στην ενότητα αυτή παρουσιάζουμε παραδείγματα της λειτουργίας της μηχανής αναζήτησης μέσω των οποίων γίνεται κατανοητή η χρησιμότητα των επεκτάσεων που πραγματοποιήθηκαν, ταυτόχρονα όμως αναδεικνύονται επιπλέον προβλήματα που εισάγει το rdf μοντέλο στην αναζήτηση με λέξεις-κλειδιά.

### *6.3.1 Περιορισμός πλήθους αποτελεσμάτων μέσω αναζήτησης στα abstracts της dbpedia*

Ερώτημα: Woody Allen, Scarlett Johansson, Film

Πριν την εφαρμογή αναζήτησης στα abstracts της DBpedia για τη λέξη κλειδί Film ανακτώνται 167875 αποτελέσματα, εκ των οποίων τα πρώτα 60253 έχουν λάβει το ίδιο σκορ λόγω κειμενικής ομοιότητας (lucene score) με τη λέξη Film:

| **Σειρά Επιστροφής** | **Resource** | **Λόγος Ανάκτησης** | **Lucene Score** |
|---|---|---|---|
| 1 | FiLm | label | 6.34617 |
| 2 | Film_%28film%29 | label | 6.34617 |
| 3 | This_Film_Is_On | label | 6.34617 |
| 4 | A-Film | label | 6.34617 |
| 5 | Film_%28television_film%29 | label | 6.34617 |
| 6 | A_Film | label | 6.34617 |
| 7 | A._Film | label | 6.34617 |
| 8 | Film | label | 6.34617 |
| 9 | /Film | label | 6.34617 |
| 10 | /film | label | 6.34617 |
| 11 | %5CFilm | label | 6.34617 |
| 12 | Actrius | ontology/Film | 6.34617 |
| | ---------- | | |
| 60250 | Love_Does_Grow_on_Trees_%28short_film%29 | Category:Film | score=6.34617 |
| 60251 | Twin_films | Category:Film | score=6.34617 |
| 60252 | Social_film | Category:Film | 6.34617 |
| 60253 | Young_Voices_on_Climate_Change | Category:Film | 6.34617 |

**Πίνακας 12: Αρχικά αποτελέσματα για τη λέξη "Film"**

Είναι προφανώς αδύνατο να αναζητηθούν μονοπάτια μεταξύ Woody Allen, Scarlett Johansson και των 60253 αποτελεσμάτων για τη λέξη film, καθώς αυτό θα έπαιρνε μη αποδεκτά πολύ χρόνο. Είναι αναγκαίο να μειώσουμε κατά πολύ το πλήθος των αποτελεσμάτων για την αναζήτηση μονοπατιών. Καθώς είναι αδύνατο να γνωρίζει η μηχανή αναζήτησης ποια αποτελέσματα να κρατήσει, θα αποκλείσει αυθαίρετα όσα βρίσκονται κάτω

από μια θέση στη λίστα αποτελεσμάτων. Με τον τρόπο αυτό, είναι εξαιρετικά πιθανό (και πράγματι έτσι συμβαίνει στο συγκεκριμένο παράδειγμα) να αποκλειστούν όλες οι σχετικές εγγραφές και να μην καταφέρει να επιστρέψει κανένα αποτέλεσμα στο χρήστη. Υπενθυμίζουμε ότι ο χρήστης βλέπει μόνο τα τελικά αποτελέσματα-γράφους, άρα δεν έχει κανέναν τρόπο να γνωρίζει γιατί η μηχανή δεν μπόρεσε να ανακτήσει αποτελέσματα.

Μετά την αναζήτηση στα abstracts και την ανάθεση νέων score στα αρχικά αποτελέσματα, ακόμα και αν περιορίσουμε στο ελάχιστο το πλήθος των αποτελεσμάτων που θέλουμε να κρατηθούν για τη λέξη Film , βλέπουμε πως πράγματι η μηχανή θα μπορέσει να φέρει στις 3 πρώτες θέσεις τις 3 ταινίες που σχετίζονται και με τους δύο αυτούς ηθοποιούς:

| Σειρά Επιστροφής | Resource | Λόγος Ανάκτησης | Score |
|---|---|---|---|
| 1 | Vicky_Cristina_Barcelona | ontology/Film | 12.69234 |
| 2 | Scoop_%282006_film%29 | ontology/Film | 12.69234 |
| 3 | Match_Point | ontology/Film | 12.69234 |

**Πίνακας 13: Τελικά αποτελέσματα για τη λέξη "Film"**

Έτσι σε δεύτερο στάδιο ανακτώνται με επιτυχία και επιστρέφονται στο χρήστη τα μονοπάτια που συνδέουν το Woody Allen, τη Scarlett Johansson και καθεμία από τις τρεις αυτές ταινίες.

Ας σημειωθεί ότι κατέστη δυνατό να φέρουμε τα 3 αυτά αποτελέσματα στις πρώτες θέσεις της λίστας, επειδή τα ονόματα και των δύο ηθοποιών αναφέρονται στα abstracts των αντίστοιχων εγγραφών. Δε θα είχαμε εξίσου καλά αποτελέσματα για παράδειγμα στην περίπτωση αναζήτησης όλων των ταινιών της Scarlett Johansson. Για το query "Scarlett Johansson, Film" ενώ στο μεγαλύτερο μέρος τους τα αποτελέσματα είναι ικανοποιητικά και ανακτώνται με επιτυχία οι περισσότερες ταινίες της ηθοποιού, συναντώνται τα εξής δύο προβλήματα:

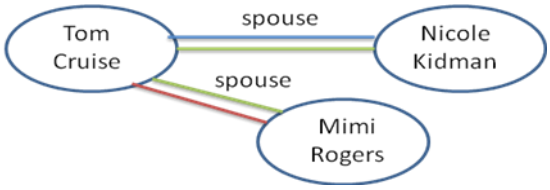
Η ταινία Prestige θα ανακτηθεί από την αναζήτηση στα abstracts, αλλά δε θα εμφανιστεί στα τελικά αποτελέσματα-μονοπάτια, επειδή η αντίστοιχη σχέση starring δεν έχει προστεθεί στα δεδομένα της dbpedia. Για τον ακριβώς αντίθετο λόγο δε θα εμφανιστεί ούτε η ταινία Ghost World στα τελικά αποτελέσματα. Ενώ η σχέση starring υπάρχει στο dataset της dbpedia, επειδή το όνομα της ηθοποιού δεν περιέχεται στο abstract της ταινίας, η αντίστοιχη εγγραφή θα αποκλειστεί από τα αποτελέσματα πριν το στάδιο της αναζήτησης μονοπατιών. Πρόκειται για προβλήματα ασυνέπειας στο dataset της dbpedia και η αντιμετώπισή τους δε θα μας απασχολήσει στην παρούσα εργασία. Ωστόσο τα αναφέρουμε γιατί πρόκειται για προβλήματα που επανεμφανίζονται και σε επόμενα ερωτήματα με αποτέλεσμα η μηχανή αναζήτησης να μην επιστρέφει πάντα τα αναμενόμενα αποτελέσματα.

### *6.3.2 Μορφή τελικών αποτελεσμάτων*

Ερώτημα : Tom Cruise, Nicole Kidman, Mimi Rogers

Σκοπός μας εδώ είναι να γίνει αντιληπτός ο τρόπος σχηματισμού των μονοπατιών-αποτελεσμάτων, η χρήση cosine similarity για το συνδυασμό τους, αλλά και να καταδειχθούν τα προβλήματα που συναντώνται. Εφόσον έχουμε 3 φράσεις-κλειδιά, κάθε επιμέρους αποτέλεσμα θα αποτελείται από 3 μονοπάτια τα οποία συνδέουν ανά δύο τις φράσεις-κλειδιά. Παρουσιάζουμε εδώ τα 7 πρώτα αποτελέσματα που επιστρέφει η μηχανή αναζήτησης για το παραπάνω ερώτημα. Κάθε ξεχωριστό μονοπάτι παρουσιάζεται με διαφορετικό χρώμα ακμών.

Το πρώτο αποτέλεσμα είναι το εξής:



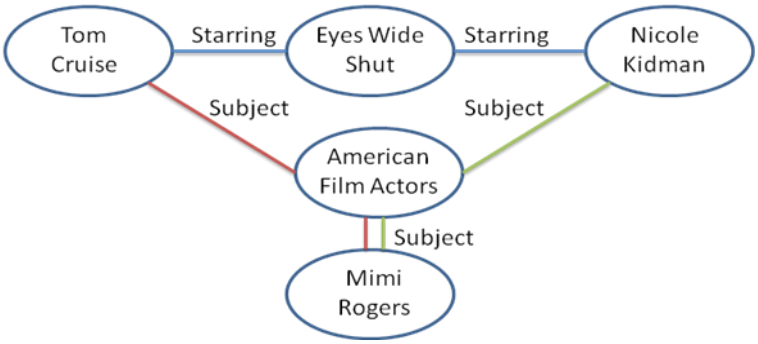

**Αποτέλεσμα 1**

Καθώς δεν περιλαμβάνεται καμία πληροφορία σχετική με το χώρο των ταινιών, τα επιμέρους μονοπάτια θα είχαν απορριφθεί στην κανονική λειτουργία της μηχανής αναζήτησης. Τα κρατήσαμε ωστόσο εδώ για να δούμε ότι πράγματι η χρήση cosine similarity ταίριαξε μεταξύ τους σωστά τις σχέσεις <spouse>, χωρίς να αναγνωρίζει βέβαια ότι πρόκειται για την ίδια σχέση. Με βάση όσα έχουμε πει είναι προφανές ότι ο παραπάνω συνδυασμός οφείλεται στην 7η διάσταση των διανυσμάτων που αντιπροσωπεύουν τα μονοπάτια, τη διάσταση που αντιστοιχεί στις μη σχετικές με τον τομέα των ταινιών σχέσεις/ακμές. Ένας άλλος λόγος που επιλέξαμε το συγκεκριμένο παράδειγμα, είναι γιατί μέσω αυτού καταδεικνύεται το πρόβλημα της επικάλυψης μονοπατιών στα αποτελέσματα εξαιτίας του τρόπου με τον οποίο η μηχανή αναζήτησης πραγματοποιεί την αναζήτηση μονοπατιών.

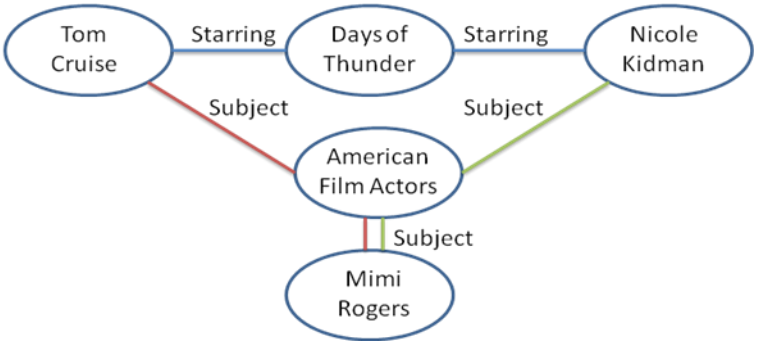
Ενώ στο συγκεκριμένο παράδειγμα η αναγνώριση της πλεονάζουσας πληροφορίας είναι εύκολη, όπως και ο εν συνεχεία αποκλεισμός της από τα αποτελέσματα, αυτό δεν ισχύει στη γενική περίπτωση. Η αναγνώριση της επανάληψης ολόκληρων μονοπατιών ή και υπομονοπατιών τους είναι σημαντικό να γίνει νωρίς στη διαδικασία της αναζήτησης ώστε να μην επιβαρύνεται το σύστημα με επιπλέον εργασίες που κοστίζουν σε χρόνο και υπολογιστικούς πόρους. Εξαιτίας όμως της ευελιξίας της μηχανής αναζήτησης στο στάδιο ανάκτησης αποτελεσμάτων από το lucene, ο αποκλεισμός δεν μπορεί να γίνει πριν το στάδιο

του συνδυασμού μονοπατιών, γιατί μέχρι τότε δεν είναι καν σαφές αν πιο χρήσιμο αποτέλεσμα είναι το μεγαλύτερο μονοπάτι ή κάποιο από τα υπομονοπάτια του.

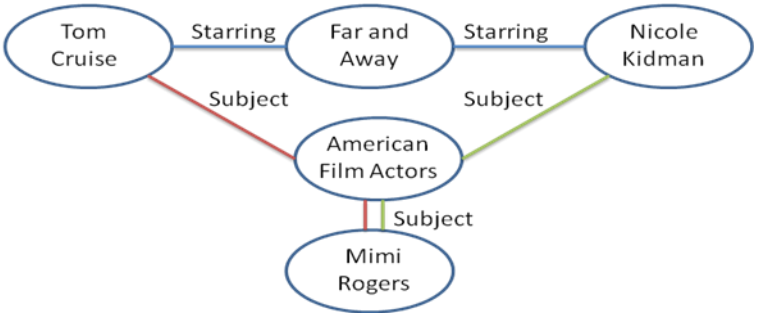
Ακολουθούν τα επόμενα 3 αποτελέσματα:




**Αποτέλεσμα 2**




**Αποτέλεσμα 3**

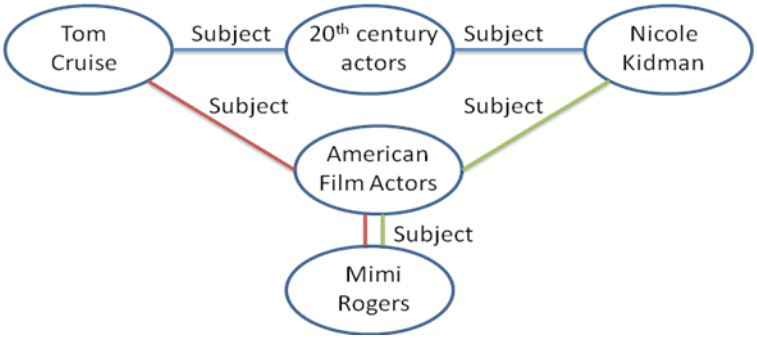



**Αποτέλεσμα 4**

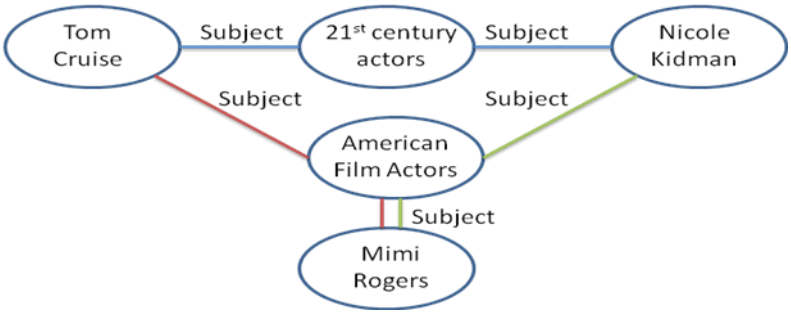
Σε καθένα από τα παραπάνω αποτελέσματα ο Tom Cruise συνδέεται με τη Nicole Kidman μέσω μιας διαφορετικής ταινίας στην οποία έπαιζαν μαζί. Καθώς η Mimi Rogers δε συνδέεται μέσω κάποιας ταινίας με τους άλλους δύο ηθοποιούς, ως πιο ταιριαστά μονοπάτια για να συγκροτηθεί το τελικό αποτέλεσμα επιλέχθηκαν από τη μηχανή αναζήτησης τα μονοπάτια εκείνα που περιλαμβάνουν πληροφορίες για τις SKOS κατηγορίες στις οποίες ανήκουν οι ηθοποιοί. Παρατηρούμε ότι τα μονοπάτια μεταξύ Mimi Rogers-Tom Cruise και Mimi Rogers-Nicole Kidman επαναλαμβάνονται ίδια και στα 3 αποτελέσματα, αφού δεν

υπάρχει εναλλακτικός τρόπος συσχέτισης της Mimi Rogers με τους άλλους δύο ηθοποιούς. Επιλέγουμε όμως να εμφανίζουμε την πληροφορία αυτή σε κάθε αποτέλεσμα για δύο λόγους. Επιθυμούμε κάθε γράφος να αποτελεί ένα πλήρες αποτέλεσμα το οποίο μπορεί να προσπελαστεί αυτόνομα από κάποιον χρήστη, χωρίς να έχουν απαραίτητα προσπελαστεί τα προηγούμενά του. Επιπλέον, με την εξατομικευμένη αναταξινόμηση των αποτελεσμάτων, δε γνωρίζουμε ποια αποτελέσματα θα βρεθούν ψηλότερα στη λίστα. Επομένως, αυτή η επανάληψη πληροφορίας στα διάφορα αποτελέσματα αποτελεί επιθυμητή λειτουργία της μηχανής αναζήτησης.

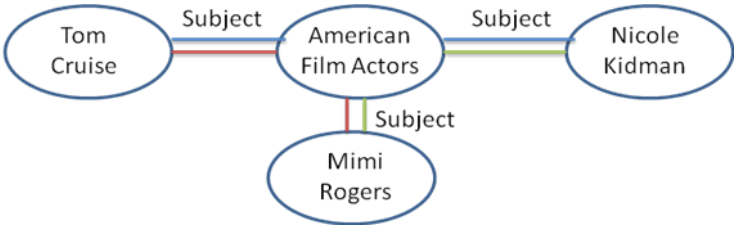
Τα επόμενα 3 αποτελέσματα:




**Αποτέλεσμα 5**




**Αποτέλεσμα 6**




**Αποτέλεσμα 7**

Σημειώνουμε πως για το συγκεκριμένο παράδειγμα, όπως και για το σύνολο των ερωτημάτων που πραγματοποιήθηκαν στα πλαίσια της παρούσας εργασίας, έχει εφαρμοστεί ο

αποκλεισμός από τα αποτελέσματα των πολύ γενικών κατηγοριών SKOS, όπως έχουμε εξηγήσει παραπάνω. Ωστόσο, ακόμα και μετά τον αποκλεισμό τους, βλέπουμε πως έχουμε στα αποτελέσματα 3 διαφορετικά μονοπάτια μεταξύ Tom Cruise-Nicole Kidman που αφορούν τις skos κατηγορίες στις οποίες ανήκουν και οι οποίες σχετίζονται με το χώρο των ταινιών. Εύκολα αντιλαμβάνεται κανείς όμως ότι η αύξηση του αριθμού των αποτελεσμάτων δε συνεπάγεται εδώ αύξηση της προσφερόμενης πληροφορίας. Το πλήθος των skos categories, η αντιστοίχιση κάθε άρθρου της Wikipedia σε περισσότερες από μια κατηγορίες και η ύπαρξη κατηγοριών που περιγράφουν έννοιες σε διαφορετικά επίπεδα γενικότητας, μπορεί να δημιουργήσει σοβαρά προβλήματα απόδοσης στη μηχανή αναζήτησης μέσω του πολλαπλασιασμού των επιστρεφόμενων αποτελεσμάτων. Είναι συνεπώς απαραίτητος ο ορισμός αυστηρών κανόνων αποδοχής ή μη μιας skos κατηγορίας ως χρήσιμου αποτελέσματος οι οποίοι θα εξαρτώνται και από την εφαρμογή που χρησιμοποιεί τη μηχανή αναζήτησης.

Τέλος, πρέπει να αναφερθεί ότι στο παράδειγμα που εξετάζουμε εδώ οι κόμβοι που αντιστοιχούν στις φράσεις κλειδιά συνδέονται ανά δύο με μονοπάτια. Αυτό φυσικά δεν είναι πάντα εφικτό, ιδιαίτερα όταν για κάθε φράση-κλειδί έχουμε κρατήσει περισσότερα από ένα αποτελέσματα. Στην περίπτωση λοιπόν που δυο κόμβοι συνδέονται μεταξύ τους αλλά όχι με τον κόμβο που αντιστοιχεί στην τρίτη λέξη-κλειδί, επιλέγουμε να εμφανίσουμε και αυτό το ημιτελές αποτέλεσμα, χαμηλότερα ίσως στη λίστα αποτελεσμάτων. Ο λόγος είναι ότι πιθανώς αυτό το αποτέλεσμα να είναι αρκετά ικανοποιητικό, σε περίπτωση ειδικά που δεν υπάρχει τελικά γράφος που να περιλαμβάνει όλες τις λέξεις-κλειδιά.

# 7

# *Πειράματα και Αξιολόγηση*

Έχοντας εμπλουτίσει το αρχικό σύνολο των δεδομένων εισόδου και έχοντας δημιουργήσει τα χαρακτηριστικά εκπαίδευσης του SVM για τα δεδομένα αυτά, στο κεφάλαιο αυτό θα αξιολογήσουμε τη χρήση του Ranking SVM στην εξατομικευμένη ταξινόμηση αποτελεσμάτων.

## *7.1 Οργάνωση Πειραμάτων*

### *7.1.1 Επιλογή χρηστών για τα πειράματα*

Θέλουμε οι χρήστες που θα χρησιμοποιηθούν στα πειράματα να ικανοποιούν τα κριτήρια επιλογής που ορίσαμε στο κεφάλαιο 3. Επιπλέον, θέλουμε το τελευταίο από τα κριτήρια, δηλαδή η σχετικά ομοιόμορφη κατανομή βαθμολογιών να ικανοποιείται ξεχωριστά στο training set και στο testing set (κρατώντας το 80% των συνολικών δεδομένων ως training set και το υπόλοιπο 20% ως testing set). Η απαίτηση αυτή δεν είναι αυστηρή, εφόσον όμως οι βαθμολογίες του χρήστη σε κάποιο από τα δύο αυτά σύνολα δεν είναι στο σύνολό τους χαμηλές ή υψηλές. Με αυτά τα κριτήρια καταλήγουμε στο εξής σύνολο 25 χρηστών το οποίο χωρίζεται στις εξής 2 κατηγορίες:

- Χρήστες με πολύ μεγάλο πλήθος βαθμολογιών

| ID χρήστη | πλήθος ταινιών | Μ.Ο. ταινιών | #ταινιών με βαθμό 5 | #ταινιών με βαθμό 4 | #ταινιών με βαθμό 3 | #ταινιών με βαθμό 2 | #ταινιών με βαθμό 1 |
|---|---|---|---|---|---|---|---|
| A1 | 2185 | 2.84 | 17 | 107 | 236 | 64 | 13 |
| A2 | 2206 | 2.68 | 10 | 75 | 129 | 149 | 78 |
| A3 | 1649 | 2.93 | 17 | 102 | 140 | 62 | 8 |
| A4 | 1770 | 2.57 | 17 | 50 | 133 | 124 | 30 |
| A5 | 1752 | 3.18 | 55 | 89 | 121 | 71 | 14 |
| A6 | 1726 | 2.91 | 27 | 71 | 105 | 85 | 57 |
| A7 | 1656 | 3.08 | 16 | 64 | 112 | 114 | 25 |
| A8 | 2092 | 3.10 | 31 | 81 | 216 | 70 | 20 |
| A9 | 2440 | 2.57 | 7 | 44 | 149 | 162 | 126 |
| A10 | 1935 | 2.88 | 29 | 49 | 143 | 103 | 63 |
| A11 | 1840 | 3.15 | 28 | 72 | 151 | 92 | 25 |
| A12 | 1675 | 3.16 | 19 | 97 | 156 | 58 | 5 |
| A13 | 1635 | 2.84 | 17 | 71 | 122 | 96 | 21 |
| A14 | 1890 | 3.04 | 69 | 66 | 87 | 111 | 45 |

**Πίνακας 14: Χρήστες με πολλές βαθμολογίες**

- Χρήστες με σχετικά μικρό πλήθος βαθμολογιών

| ID χρήστη | πλήθος ταινιών | Μ.Ο. ταινιών | #ταινιών με βαθμό 5 | #ταινιών με βαθμό 4 | #ταινιών με βαθμό 3 | #ταινιών με βαθμό 2 | #ταινιών με βαθμό 1 |
|---|---|---|---|---|---|---|---|
| B1 | 200 | 3.61 | 3.00 | 8 | 25 | 2 | 2 |
| B2 | 167 | 3.23 | 0.00 | 11 | 12 | 7 | 3 |
| B3 | 164 | 3.48 | 5.00 | 10 | 7 | 6 | 4 |
| B4 | 180 | 3.29 | 0.00 | 12 | 12 | 8 | 4 |
| B5 | 177 | 2.42 | 0.00 | 6 | 12 | 8 | 9 |
| B6 | 264 | 3.13 | 0.00 | 8 | 43 | 1 | 0 |
| B7 | 154 | 3.06 | 3.00 | 6 | 12 | 6 | 3 |
| B8 | 174 | 3.07 | 2.00 | 14 | 10 | 7 | 1 |
| B9 | 291 | 3.05 | 0.00 | 11 | 40 | 7 | 0 |
| B10 | 513 | 3.63 | 6.00 | 31 | 59 | 6 | 0 |
| B11 | 151 | 3.20 | 2.00 | 9 | 15 | 4 | 0 |

**Πίνακας 15: Χρήστες με λίγες βαθμολογίες**

Στους παραπάνω πίνακες τα μεγέθη αναφέρονται στο σύνολο των διαθέσιμων ταινιών κάθε χρήστη, δηλαδή σε αυτές που θα χρησιμοποιηθούν στην εκπαίδευση και στην αξιολόγηση.

### *7.1.2 Ποσοστό δεδομένων που χρησιμοποιούνται στην εκπαίδευση*

Στα πειράματά μας ένα ποσοστό των συνολικών δεδομένων για κάθε χρήστη χρησιμοποιείται για την εκπαίδευση του SVM (training set) και τα υπόλοιπα χρησιμοποιούνται για την μέτρηση της απόδοσής του (testing set). Για την εύρεση του κατάλληλου ποσοστού δοκιμάσαμε τις τιμές 20%, 40%, 60%, 80% και 90%. Βρήκαμε ότι καταλληλότερη είναι η χρήση του 80% των δεδομένων ως training set και του υπόλοιπου 20% ως testing set. Με τον τρόπο αυτό εξασφαλίζεται η επαρκής εκπαίδευση του συστήματος, αλλά και η διατήρηση επαρκών δεδομένων για την αξιολόγησή του.

## *7.2 Πειράματα και αποτελέσματα*

### *7.2.1 Αξιολόγηση εξατομικευμένης ταξινόμησης ταινιών*

Μετά την εκπαίδευση του SVM θέλουμε να αξιολογήσουμε αρχικά την ικανότητά του να αναταξινομεί τις ταινίες του χρήστη ώστε εκείνες που ανταποκρίνονται περισσότερο στις προτιμήσεις του να έρχονται πιο ψηλά στη λίστα. Δεδομένου ότι οι ταινίες αποτελούν το βασικό στοιχείο του συνόλου των δεδομένων μας και μέσω των βαθμολογιών τους έγινε η εκπαίδευση του SVM, αυτό θα είναι το κύριο μέτρο αξιολόγησης της αποτελεσματικότητας της μεθόδου.

Τα διαγράμματα 1-25 αφορούν εκπαίδευση του SVM με χρήση του 80% των βαθμολογιών κάθε χρήστη ως training set. Στα συγκεκριμένα διαγράμματα η εκπαίδευση έχει γίνει με χρήση και των 10 ομάδων χαρακτηριστικών που περιγράφηκαν στο κεφάλαιο 5. Για την αξιολόγηση χρησιμοποιήθηκε το υπόλοιπο 20% των βαθμολογιών. Ο διαχωρισμός των δεδομένων σε training και testing set έγινε με βάση τις ημερομηνίες που ο χρήστης έδωσε τις βαθμολογίες, έτσι ώστε τα στοιχεία του training να προηγούνται χρονικά των στοιχείων του testing.

Καθεμία από τις 3 καμπύλες ενός διαγράμματος αφορά μια διαφορετική ταξινόμηση όλων των ταινιών του testing set. Το σημείο (x,y) μιας καμπύλης αντιστοιχεί στο μέσο όρο y των ταινιών που βρίσκονται στις πρώτες x θέσεις της λίστας ταινιών όπως αυτή έχει διαμορφωθεί για καθεμία από τις διαφορετικές ταξινομήσεις.

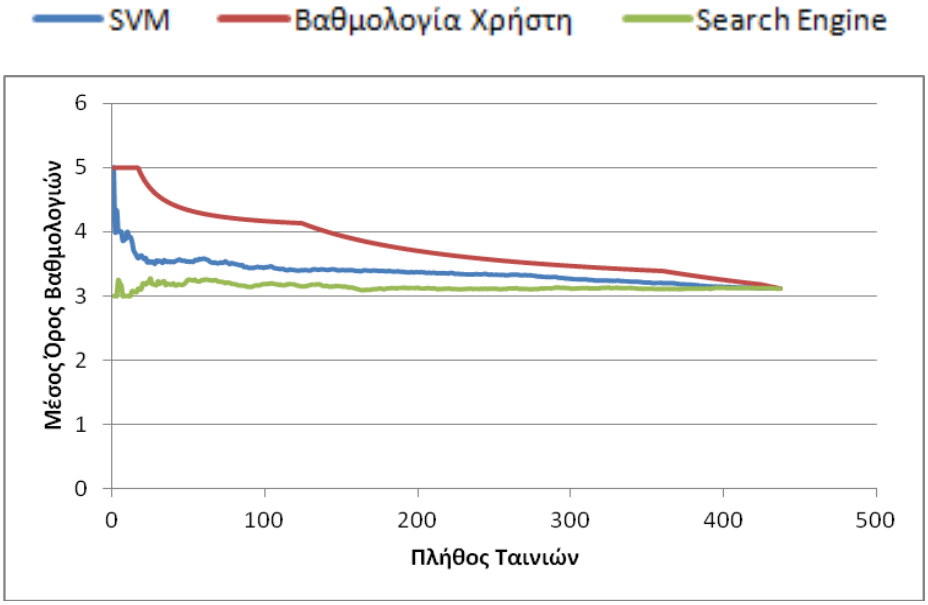
Η κόκκινη γραμμή (Βαθμολογία Χρήστη) αντιστοιχεί στη βέλτιστη δυνατή ταξινόμηση, αυτή δηλαδή βασίζεται στις πραγματικές βαθμολογίες που έχουν δοθεί από το χρήστη.

Η πράσινη γραμμή (Search Engine) αντιστοιχεί στη σειρά επιστροφής των ταινιών από τη μηχανή αναζήτησης, δηλαδή στο lucene score των αποτελεσμάτων που ανακτά η μηχανή για τη λέξη-κλειδί "Film".

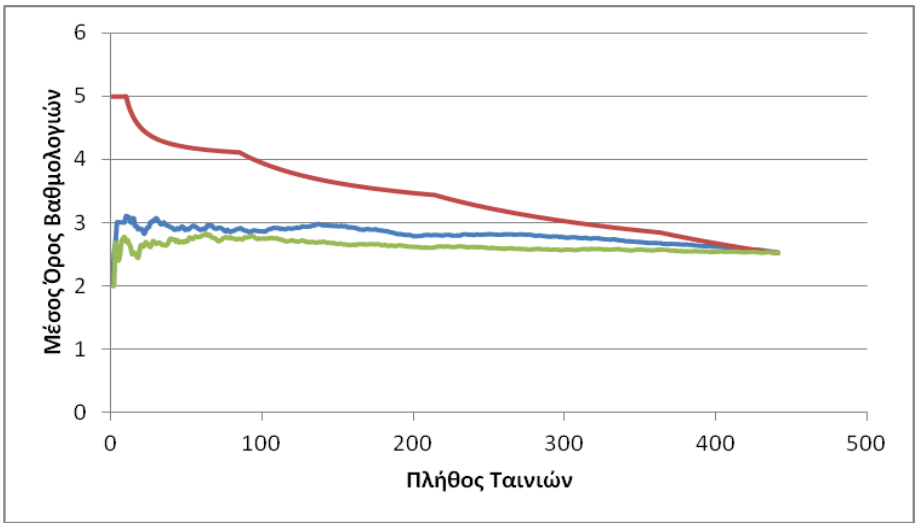
Τέλος, η μπλε γραμμή (SVM) αντιστοιχεί στην ταξινόμηση των ταινιών από το Ranking SVM.

Τα διαγράμματα δίνονται ξεχωριστά για τους χρήστες καθεμίας από τις δύο κατηγορίες χρηστών που ορίστηκαν στην παράγραφο 7.1.1.

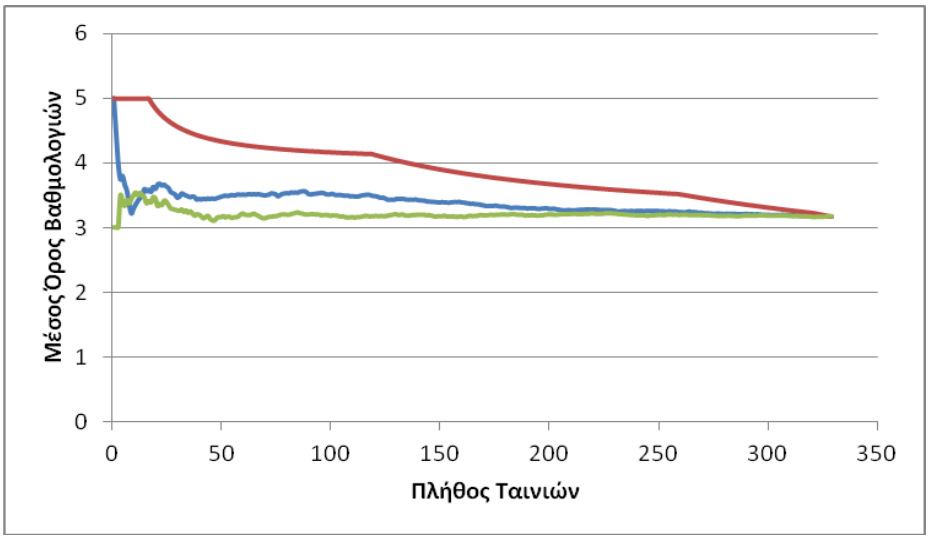
#### *7.2.1.1 Χρήστες με μεγάλο πλήθος βαθμολογιών*



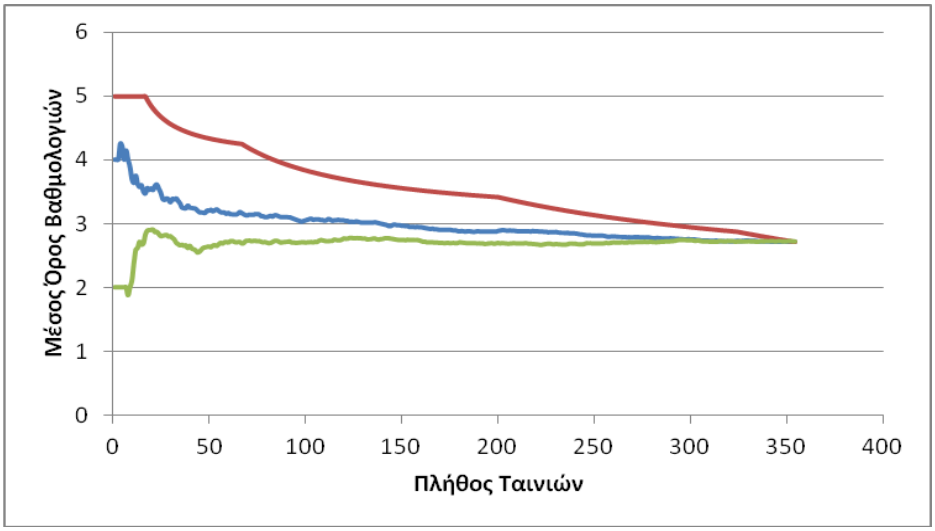

**Διάγραμμα 1: rerank ταινιών για το χρήστη Α1**



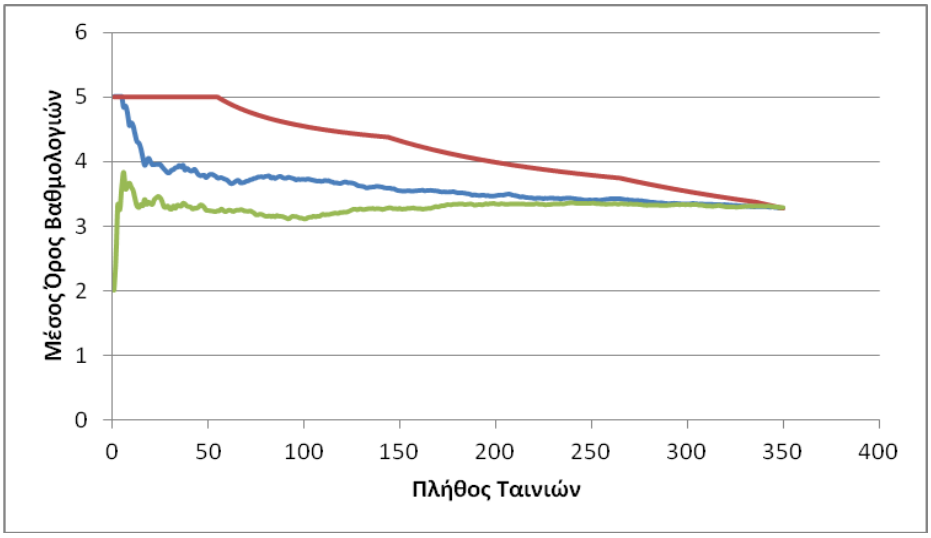

**Διάγραμμα 2: rerank ταινιών για το χρήστη Α2**



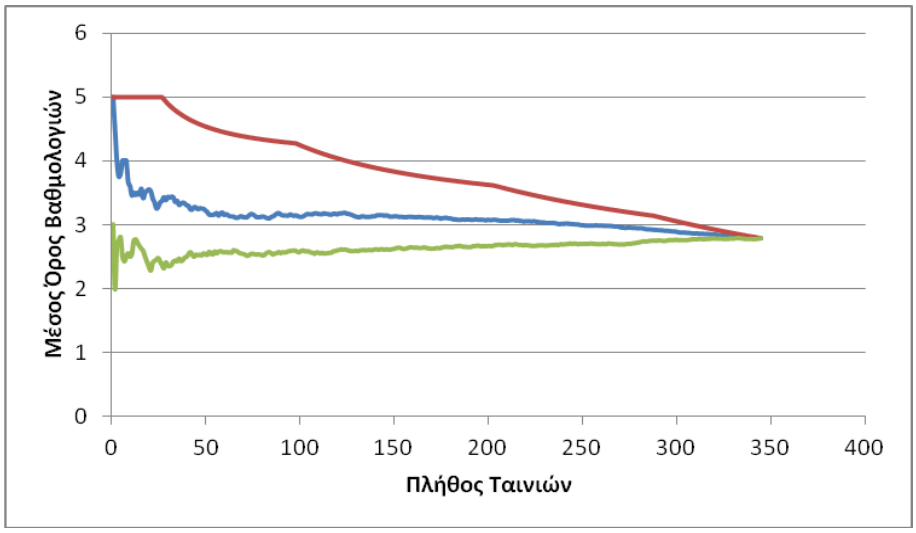

**Διάγραμμα 3: rerank ταινιών για το χρήστη Α3**



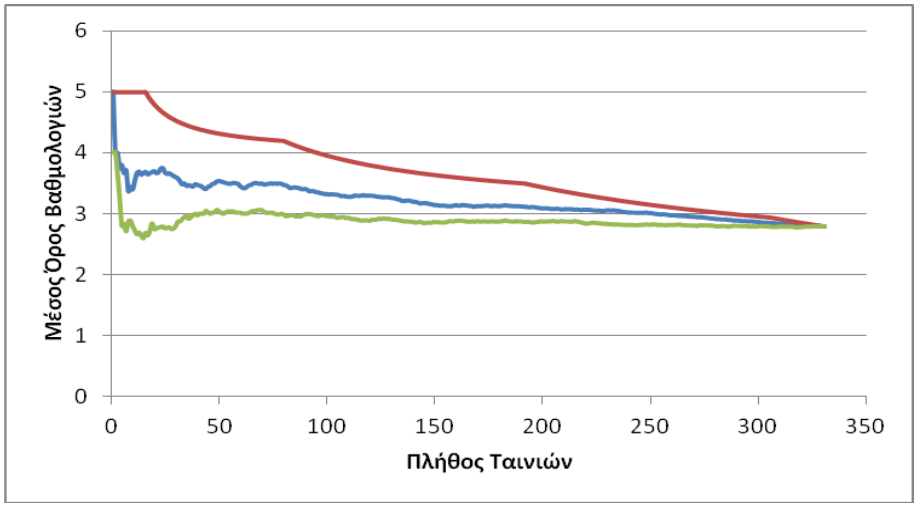

**Διάγραμμα 4: rerank ταινιών για το χρήστη Α4**



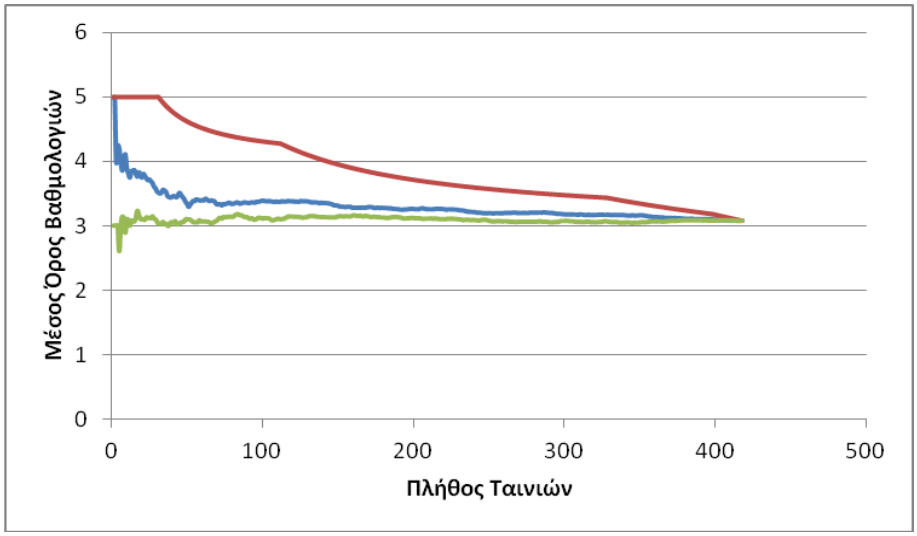

**Διάγραμμα 5: rerank ταινιών για το χρήστη Α5**



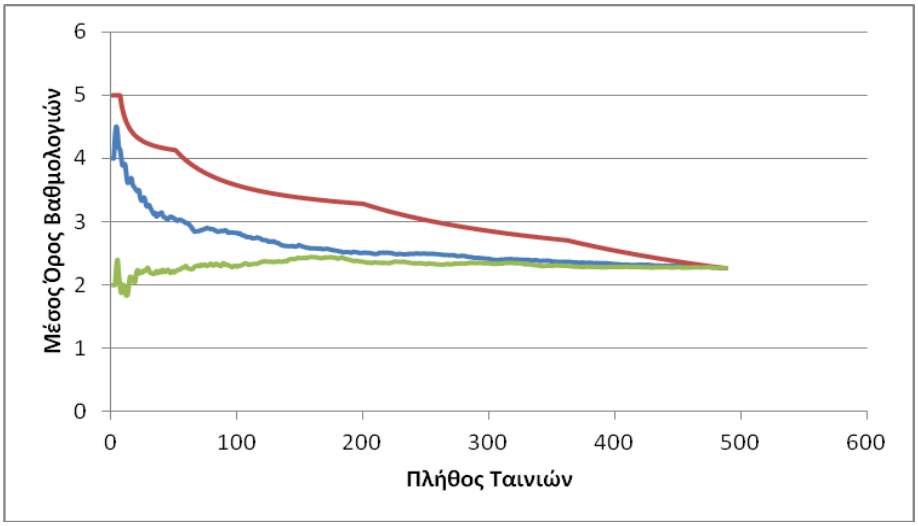

**Διάγραμμα 6: rerank ταινιών για το χρήστη Α6**



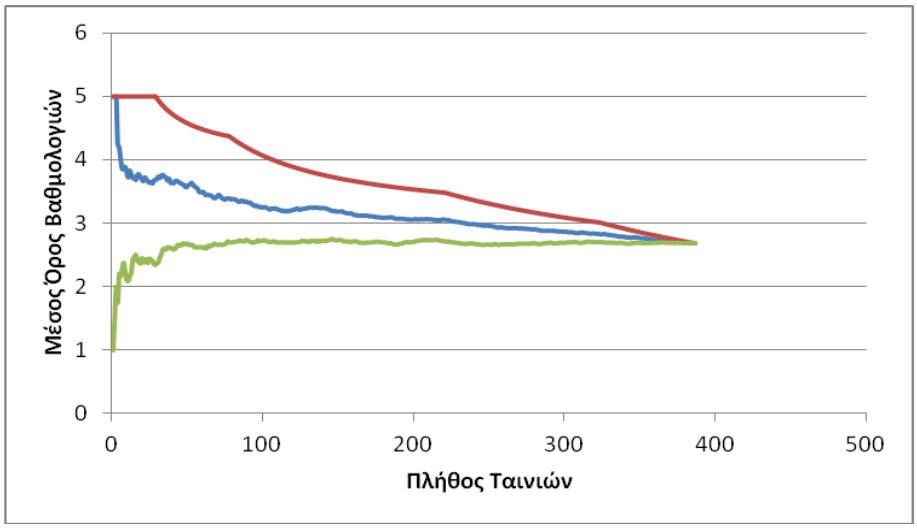

**Διάγραμμα 7: rerank ταινιών για το χρήστη Α7**



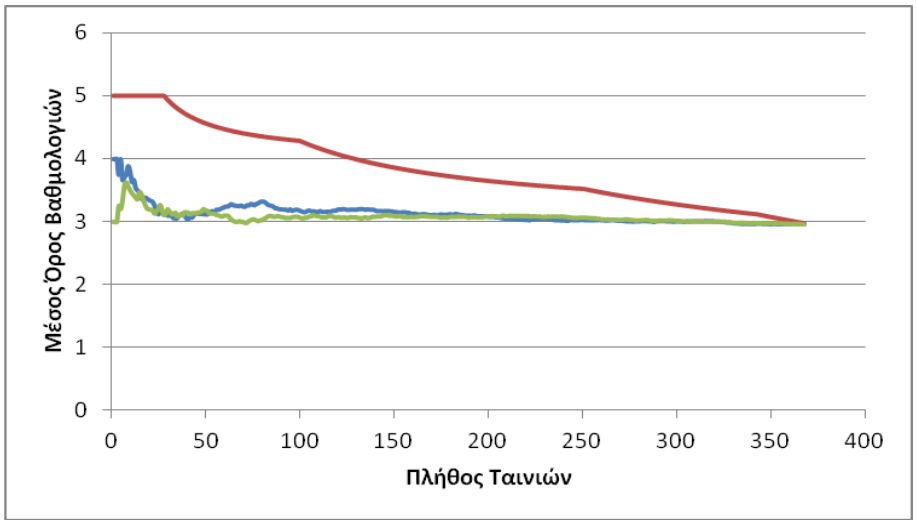

**Διάγραμμα 8: rerank ταινιών για το χρήστη Α8**



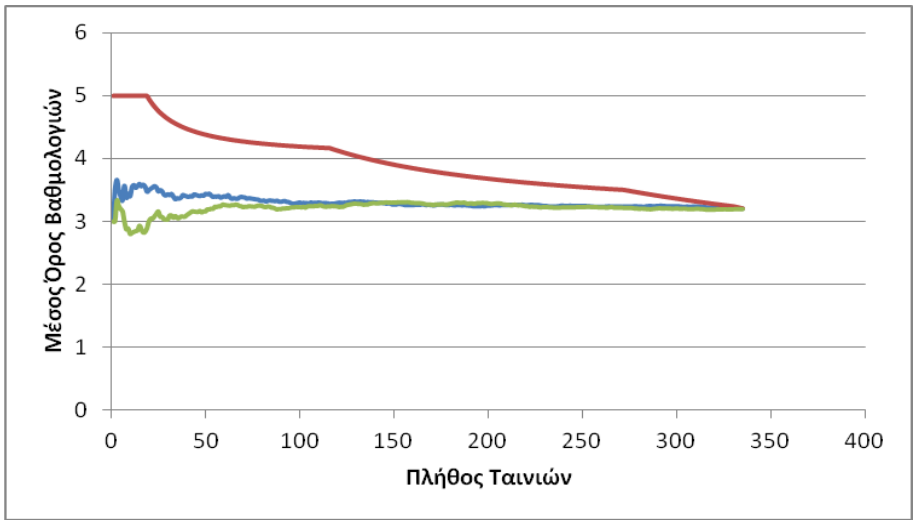

**Διάγραμμα 9: rerank ταινιών για το χρήστη A1**



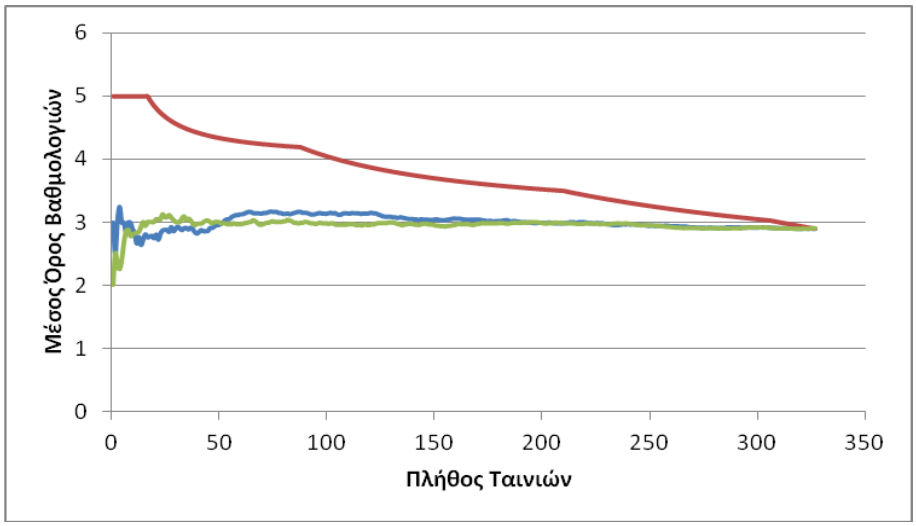

**Διάγραμμα 10: rerank ταινιών για το χρήστη A10**



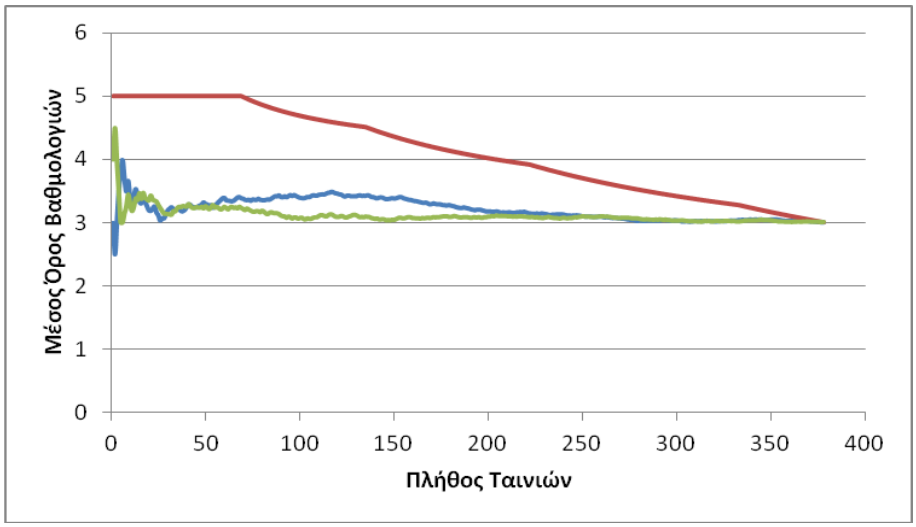

**Διάγραμμα 11: rerank ταινιών για το χρήστη A11**




**Διάγραμμα 12: rerank ταινιών για το χρήστη A12**




**Διάγραμμα 13: rerank ταινιών για το χρήστη A13**



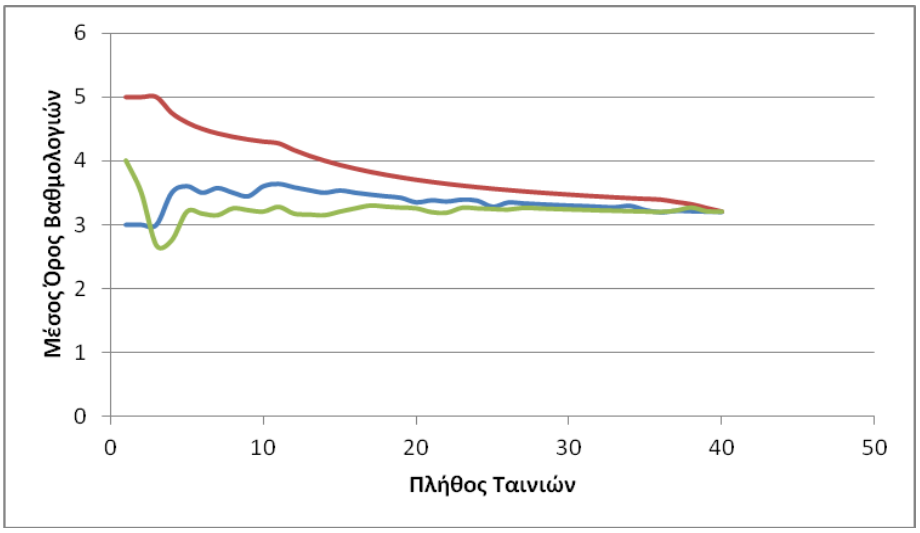

**Διάγραμμα 14: rerank ταινιών για το χρήστη A14**

**Παρατηρήσεις**

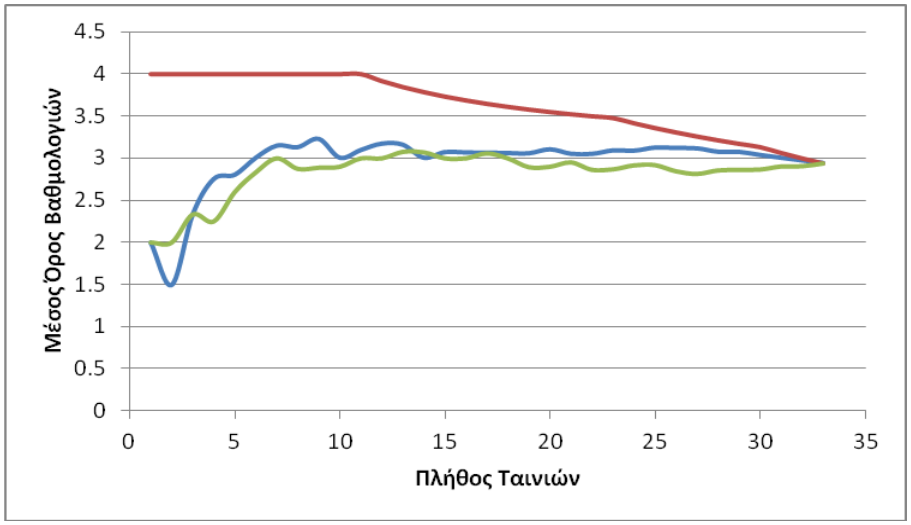
- Η πράσινη γραμμή κινείται πολύ κοντά στο μέσο όρο των συνολικών βαθμολογιών. Αυτό είναι απόλυτα φυσιολογικό, αν αναλογιστούμε ότι όλα τα αποτελέσματα που ανακτώνται σε αυτό το στάδιο από τη μηχανή αναζήτησης έχουν το ίδιο lucene score, αφού έχουν όλα ανακτηθεί λόγω της κειμενικής ομοιότητας της κλάσης τους (Film) με τη λέξη κλειδί film. Συνεπώς η σειρά επιστροφής είναι ουσιαστικά τυχαία και η καμπύλη που της αντιστοιχεί δε θα μπορούσε να είναι σημαντικά μετατοπισμένη πάνω ή κάτω του συνολικού μέσου όρου των βαθμολογιών του χρήστη.
- Σκοπός μας είναι να δούμε αν για μια τόσο αόριστη λέξη-κλειδί για την οποία ανακτάται τεράστιος αριθμός αποτελεσμάτων μπορεί η χρήση του SVM να αναταξινομήσει τα αποτελέσματα ώστε να ικανοποιούν περισσότερο τις προτιμήσεις του χρήστη. Πράγματι, βλέπουμε πως στην πλειοψηφία των χρηστών η ταξινόμηση των ταινιών από το SVM συμβάλλει σε αυτή την κατεύθυνση.
- Χρησιμοποιούμε ως κριτήριο την αύξηση του μέσου όρου των ταινιών που εμφανίζονται στις πρώτες x θέσεις της λίστας αποτελεσμάτων. Είναι λογικό πως από ένα σημείο x και μετά και η μπλε καμπύλη συγκλίνει στο μέσο όρο του χρήστη. Ωστόσο αυτό που μας ενδιαφέρει είναι τα αποτελέσματα που βρίσκονται σχετικά ψηλά στη λίστα να έχουν βαθμολογίες μεγαλύτερες του μέσου όρου του χρήστη. Πράγματι, η μπλε καμπύλη είναι στα περισσότερα διαγράμματα φθίνουσα, άρα στις πρώτες θέσεις της λίστας έχουν τοποθετηθεί ταινίες με μεγαλύτερες βαθμολογίες.
- Η κόκκινη καμπύλη αντιπροσωπεύει την ιδεατή ταξινόμηση των ταινιών με βάση τις πραγματικές βαθμολογίες του χρήστη και έχει συμπεριληφθεί για λόγους πληρότητας και όχι ως πραγματικό μέτρο αξιολόγησης της εξατομικευμένης ταξινόμησης του SVM. Σκοπός μας εδώ είναι να μετακινήσουμε την πράσινη καμπύλη προς τα πάνω, να βελτιώσουμε δηλαδή τη σειρά επιστροφής των ταινιών από τη μηχανή αναζήτησης με κριτήριο τις προτιμήσεις του χρήστη.

#### *7.2.1.2 Χρήστες με σχετικά μικρό πλήθος βαθμολογιών*

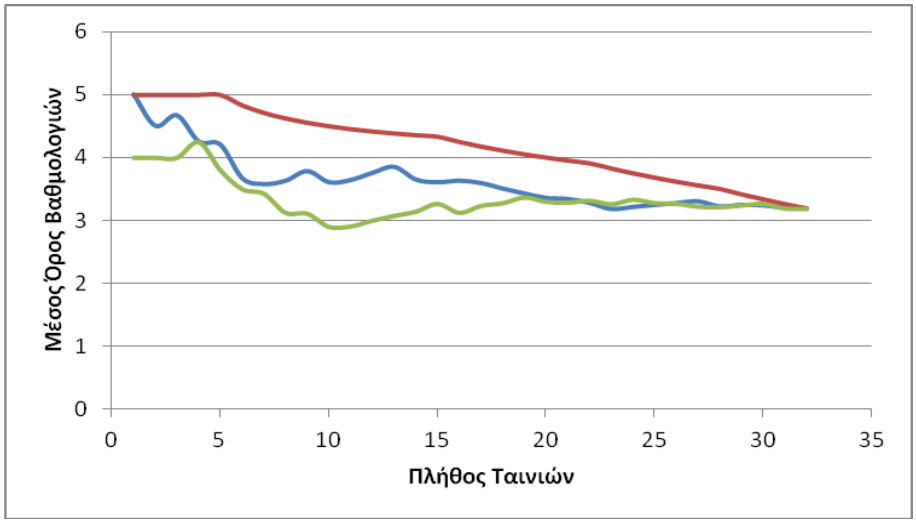
Αντίστοιχες μετρήσεις πραγματοποιήσαμε και για χρήστες με σημαντικά μικρότερο αριθμό ταινιών με στόχο να εξετάσουμε την επίδραση του μεγέθους του training set στην απόδοση του Ranking SVM. Ακολουθούν διαγράμματα για την αξιολόγηση της εξατομικευμένης ταξινόμησης ταινιών για καθέναν από τους 11 χρήστες αυτής της κατηγορίας.



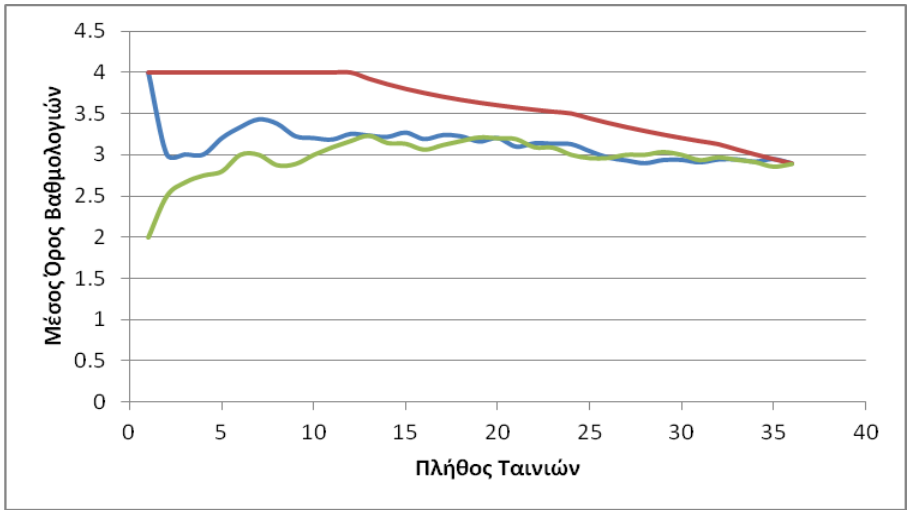

**Διάγραμμα 15: rerank ταινιών για το χρήστη Β1**



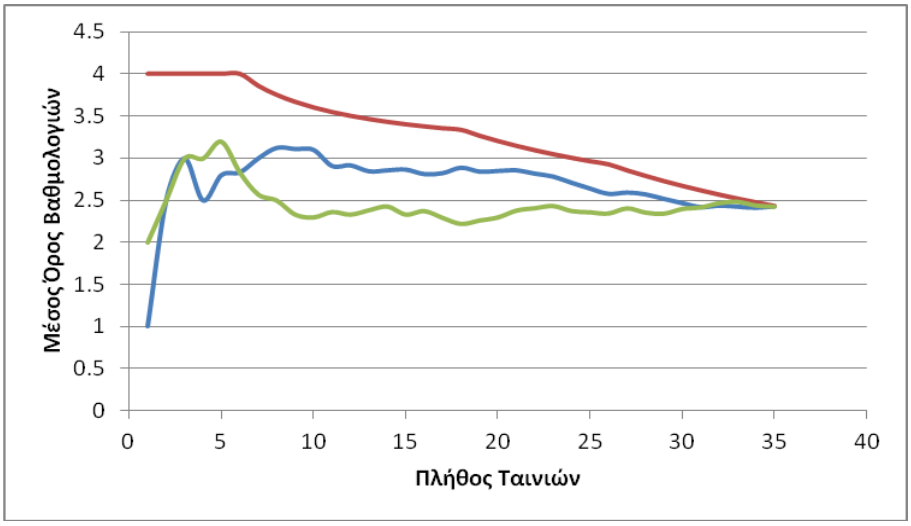

**Διάγραμμα 16: rerank ταινιών για το χρήστη Β2**



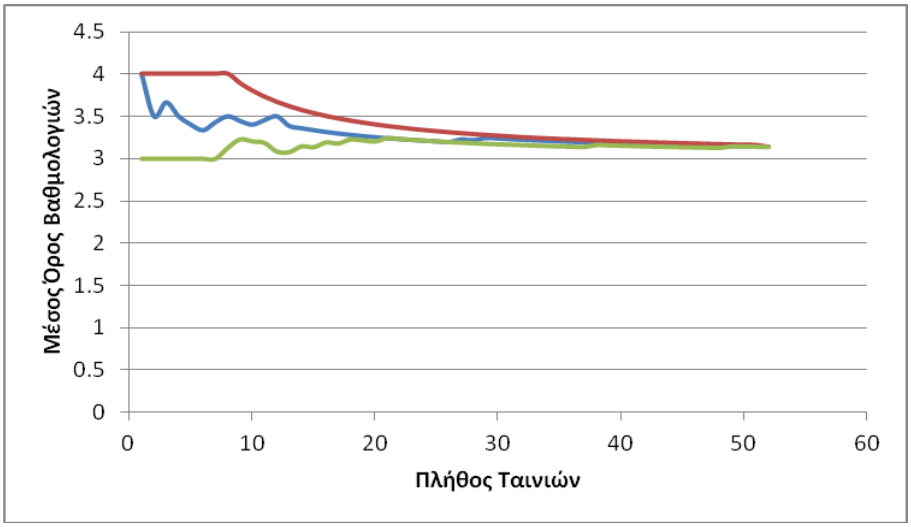

**Διάγραμμα 17: rerank ταινιών για το χρήστη Β3**



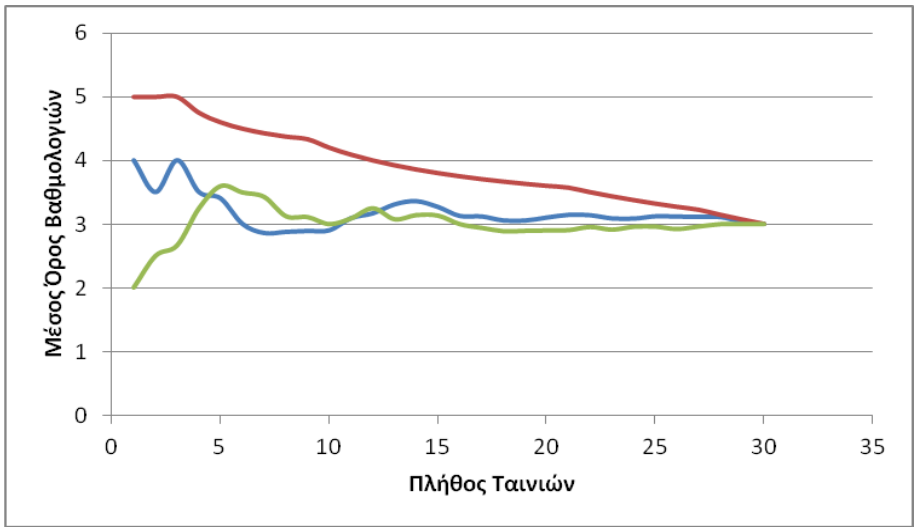

**Διάγραμμα 18: rerank ταινιών για το χρήστη Β4**



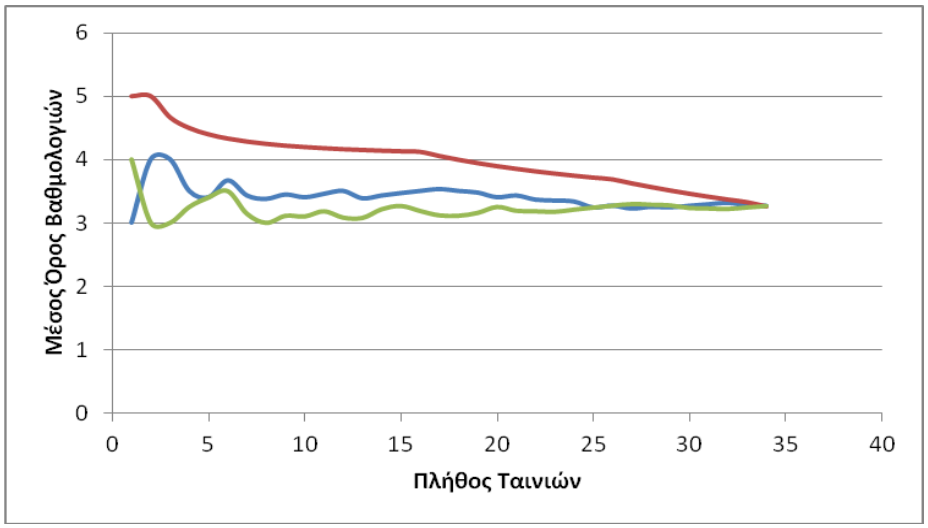

**Διάγραμμα 19: rerank ταινιών για το χρήστη Β5**



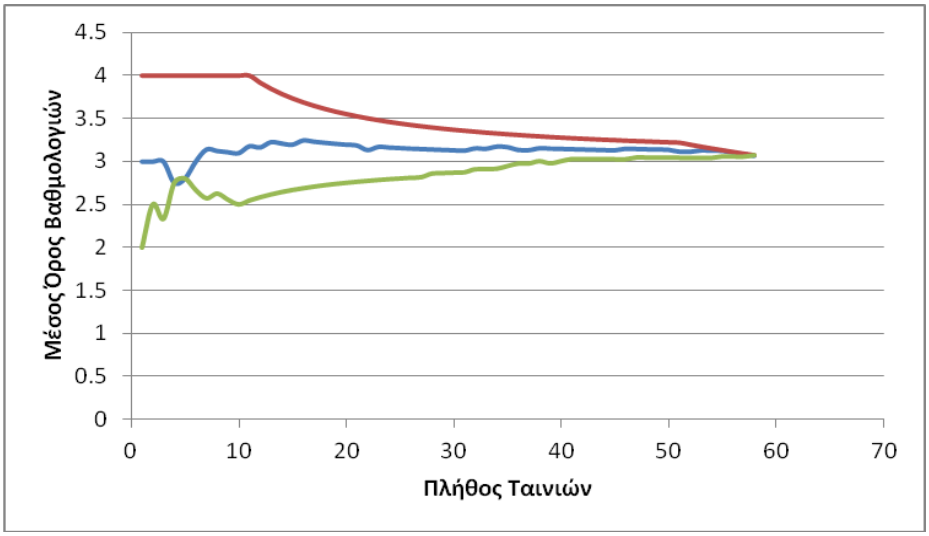

**Διάγραμμα 20: rerank ταινιών για το χρήστη Β6**



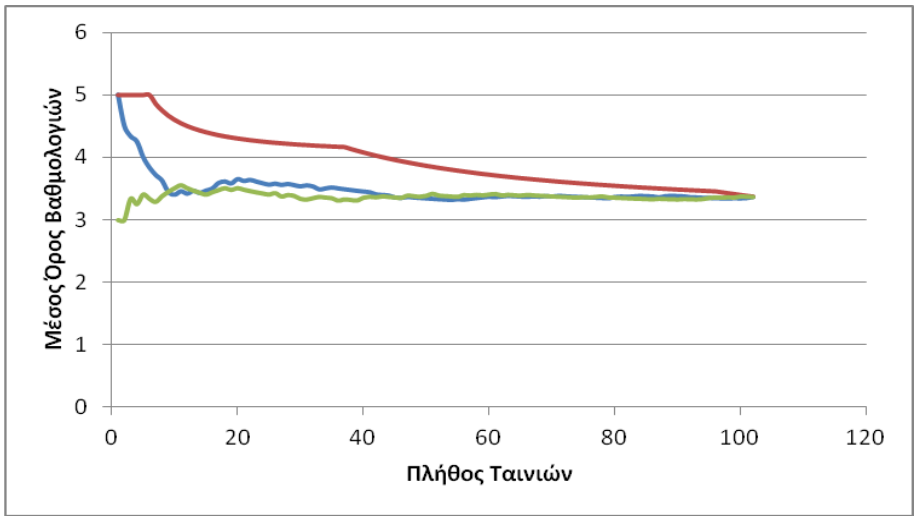

**Διάγραμμα 21: rerank ταινιών για το χρήστη Β7**



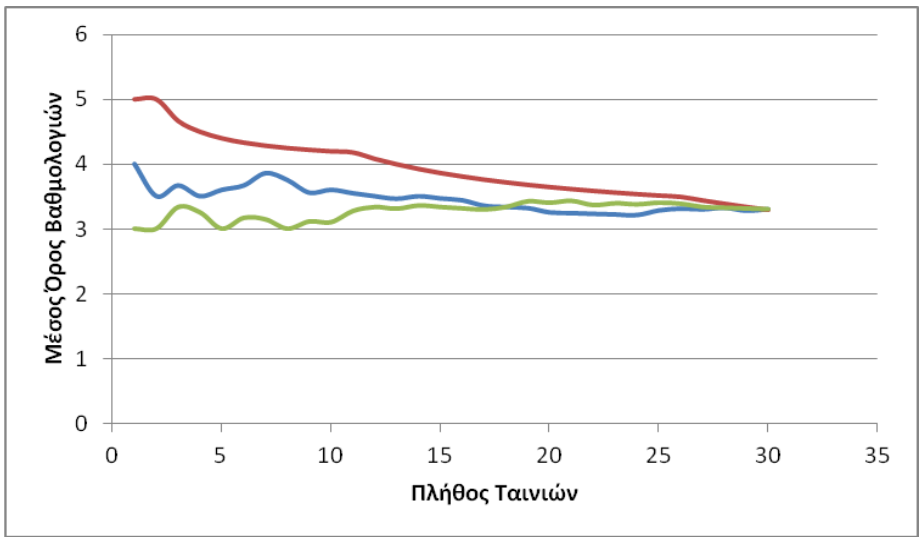

**Διάγραμμα 22: rerank ταινιών για το χρήστη Β8**




**Διάγραμμα 23: rerank ταινιών για το χρήστη Β9**




**Διάγραμμα 24: rerank ταινιών για το χρήστη Β10**




**Διάγραμμα 25: rerank ταινιών για το χρήστη Β11**

**Παρατηρήσεις**

- Και εδώ η πράσινη γραμμή κινείται, όπως περιμένουμε, κοντά στο συνολικό μέσο όρο των βαθμολογιών, ενώ η κόκκινη γραμμή αντιπροσωπεύει όπως είπαμε την ιδεατή ταξινόμηση και δεν αποτελεί πραγματικό μέτρο αξιολόγησης
- Εξαιτίας του σαφώς μικρότερου πλήθους ταινιών, βλέπουμε πως οι καμπύλες είναι πιο ευαίσθητες στις εναλλαγές βαθμολογιών.
- Και εδώ στην πλειοψηφία των χρηστών η μπλε καμπύλη που αντιστοιχεί στην ταξινόμηση από το SVM κινείται πιο ψηλά από την πράσινη, τουλάχιστον στις πρώτες θέσεις της λίστας αποτελεσμάτων.
- Φαίνεται ωστόσο πως το πλήθος των δεδομένων δεν επαρκεί για το σχηματισμό σαφούς εικόνας για την απόδοση της μεθόδου. Η μπλε καμπύλη παρουσιάζει και εδώ

μια φθίνουσα τάση, σίγουρα όμως τα αποτελέσματα δεν είναι τόσο ξεκάθαρα όσο για τους χρήστες με μεγάλα training και testing set.

- Τέλος, παρατηρούμε πως τα χειρότερα αποτελέσματα εμφανίζονται στους χρήστες αυτής της κατηγορίας που έχουν τις λιγότερες βαθμολογίες (διαγράμματα 15,16,19 και 22), γεγονός που υπογραμμίζει τη σημασία της χρησιμοποίησης επαρκών δεδομένων κατά την εκπαίδευση του SVM. Υπενθυμίζουμε ότι το ποσοστό των δεδομένω κάθε χρήστη που χρησιμοποιούνται για training και testing είναι σταθερό, συνεπώς μικρότερο testing set συνεπάγεται και μικρότερο training set.

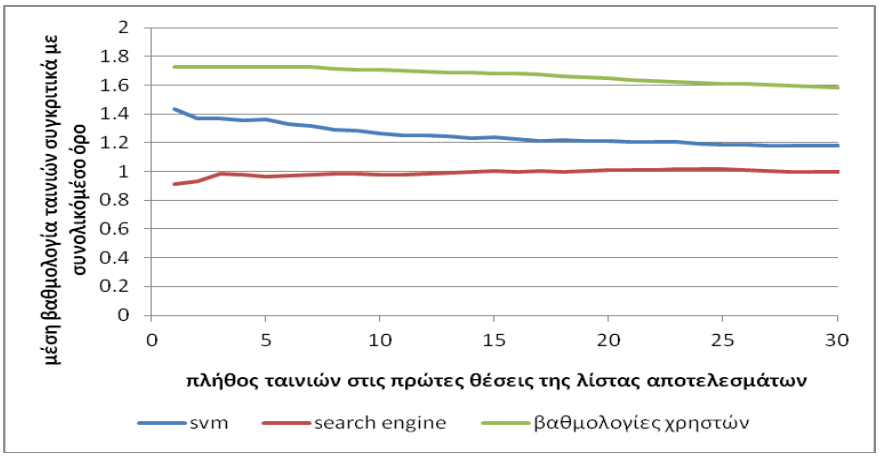
Εκτός από την προφανή επίδραση του πλήθους των βαθμολογιών που διαθέτουμε για κάθε χρήστη στην απόδοση της μεθόδου, στο κεφάλαιο 8 προτείνονται για μελλοντική μελέτη κάποιοι επιπλέον παράγοντες που πιθανώς επηρεάζουν τ το βαθμό αποτελεσματικότητας της μεθόδου για τους διάφορους χρήστες.

#### *7.2.1.3 Συγκεντρωτικά στοιχεία αξιολόγησης εξατομικευμένης ταξινόμησης ταινιών*

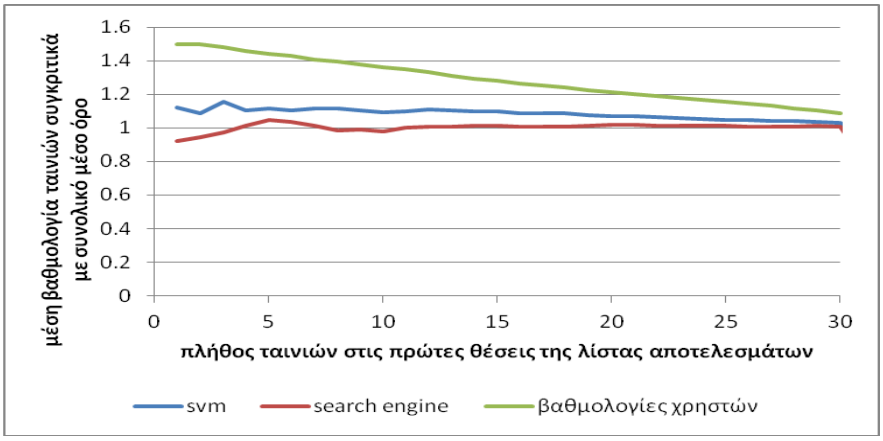
Παρουσιάζουμε εδώ 3 διαγράμματα που συνοψίζουν τα αποτελέσματα της χρήσης της μεθόδου στην εξατομικευμένη ταξινόμηση ταινιών. Το Διάγραμμα 26 αφορά τους χρήστες A1-A14, δηλαδή εκείνους για τους οποίες διαθέτουμε ένα μεγάλο πλήθος βαθμολογημένων ταινιών, το Διάγραμμα 27 αφορά τους χρήστες B1-B11 για τους οποίες διαθέτουμε σημαντικά μικρότερο πλήθος βαθμολογιών, ενώ τέλος στο Διάγραμμα 28 παρουσιάζονται τα συνολικά αποτελέσματα για τους χρήστες και των δύο ομάδων.

Ο οριζόντιος άξονας αντιστοιχεί και εδώ στο πλήθος των ταινιών που τοποθετούνται στις πρώτες x θέσεις της λίστας αποτελεσμάτων. Σταματάμε στη θέση 30, αφού για κάποιους χρήστες της δεύτερης κατηγορίας δε διαθέτουμε μεγαλύτερο αριθμό βαθμολογιών. Άλλωστε οι πρώτες θέσεις της λίστας αποτελεσμάτων είναι οι πλέον σημαντικές για την αξιολόγηση ενός συστήματος εξατομίκευσης αναζήτησης. Ο κατακόρυφος άξονας αντιστοιχεί στη βελτίωση του μέσου όρου των ταινιών αυτών συγκριτικά με το συνολικό μέσο όρο όλων των βαθμολογιών. Τα στοιχεία είναι κανονικοποιημένα ώστε η καμπύλη να μην επηρεάζεται περισσότερο από χρήστες με μεγαλύτερο μέσο όρο ή περισσότερες βαθμολογίες.

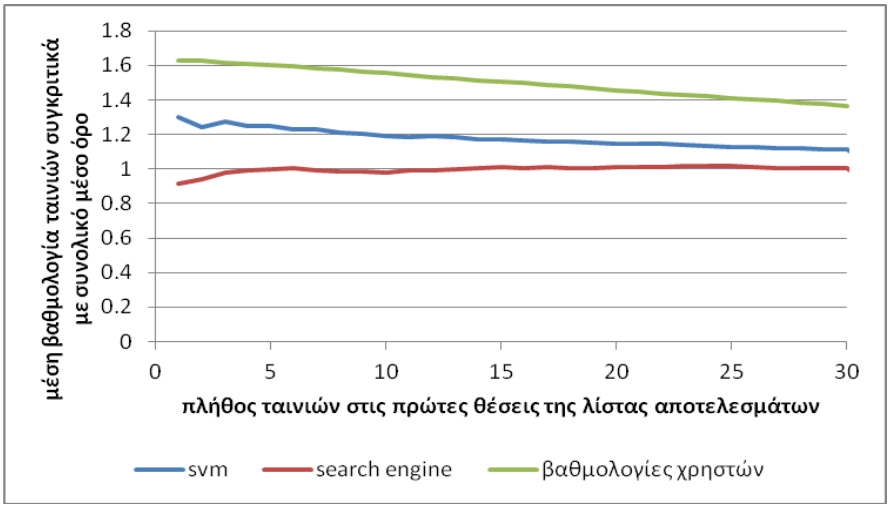
Από τα τρία διαγράμματα είναι εμφανής η συνολικά θετική επίδραση της εξατομικευμένης αναταξινόμησης ταινιών, αφού και στα τρία η καμπύλη που αντιστοιχεί στο SVM κινείται πάνω από την αντίστοιχη καμπύλη της ταξινόμησης από τη μηχανή αναζήτησης. Βλέπουμε και εδώ τη σημαντική επίδραση του πλήθους των βαθμολογιών στην αποτελεσματικότητα της μεθόδου, όπως παρατηρήθηκε και από τα μεμονωμένα διαγράμματα των χρηστών.




**Διάγραμμα 26: συγκεντρωτικά αποτελέσματα εξατομικευμένης ταξινόμησης ταινιών για τους χρήστες A1-A14**




**Διάγραμμα 27: συγκεντρωτικά αποτελέσματα εξατομικευμένης ταξινόμησης ταινιών για τους χρήστες B1-B11**




**Διάγραμμα 28: συγκεντρωτικά αποτελέσματα εξατομικευμένης ταξινόμησης ταινιών για όλους τους χρήστες**

Συνολικά η ταξινόμηση που δίνει η μέθοδός μας υπερέχει της αρχικής ταξινόμησης της μηχανής αναζήτησης και μάλιστα σε πολλές περιπτώσεις πλησιάζει ικανοποιητικά τη μορφή της ιδεατής καμπύλης και βελτιώνει σημαντικά το μέσο όρο των ταινιών που τοποθετούνται στις πρώτες θέσεις της λίστας, με τα καλύτερα αποτελέσματα να παρουσιάζονται για χρήστες με μεγάλο πλήθος βαθμολογημένων ταινιών, δηλαδή με μεγαλύτερο δεδομένων εκπαίδευσης.

### *7.2.2 Αξιολόγηση εξατομικευμένης ταξινόμησης συντελεστών ταινιών*

Στο σημείο αυτό θέλουμε να δούμε την απόδοση του Ranking SVM στην εξατομικευμένη ταξινόμηση συντελεστών ταινιών, δηλαδή τους ηθοποιούς και τους σκηνοθέτες των ταινιών που έχει βαθμολογήσει κάθε χρήστης. Υπενθυμίζουμε ότι για ηθοποιούς και σκηνοθέτες δε διαθέτουμε δεδομένα εκπαίδευσης. Η αναταξινόμηση εδώ θα βασιστεί στην εκπαίδευση του SVM για τις ταινίες του χρήστη η οποία έγινε με χαρακτηριστικά κατάλληλα επιλεγμένα ώστε να εφαρμόζονται και για ηθοποιούς και σκηνοθέτες.

Καθώς δεν είμαστε σε θέση να γνωρίζουμε τις βαθμολογίες που θα έδινε ο χρήστης για καθέναν από τους συντελεστές των ταινιών, για την αξιολόγηση της μεθόδου θα χρησιμοποιηθούν οι εξής δύο τρόποι καθορισμού της βαθμολογίας που θα έδινε ο χρήστης σε έναν ηθοποιό/σκηνοθέτη:

- ο μέσος όρος των βαθμολογιών των ταινιών στις οποίες συμμετείχε
- το συνολικό άθροισμα των βαθμολογιών των ταινιών στις οποίες συμμετείχε

Από τους δύο αυτούς τρόπους, ο πρώτος είναι μια προφανής, μάλλον, επιλογή. Ο δεύτερος τρόπος επιχειρεί να αποτυπώσει το ενδιαφέρον του χρήστη για έναν ηθοποιό όπως αυτό εκφράζεται μέσω όχι απαραίτητα πάντα πολύ υψηλών βαθμολογιών στις ταινίες του, αλλά τη συχνή επιλογή ταινιών στις οποίες συμμετέχει. Με τα κριτήρια αυτά επιχειρούμε να προσεγγίσουμε/προβλέψουμε την ταξινόμηση των ηθοποιών και σκηνοθετών που θα έδινε ο χρήστης, αφού αυτή η πληροφορία δεν είναι διαθέσιμη. Τα προβλήματα που δημιουργούνται από τους πλασματικούς αυτούς τρόπους αξιολόγησης αναλύονται παρακάτω.

Όλοι οι ηθοποιοί και οι σκηνοθέτες των ταινιών που περιλαμβάνονται στο training set και στο testing set με τις ταινίες κάθε χρήστη έχουν χρησιμοποιηθεί εδώ για την αξιολόγηση της μεθόδου. Εφόσον ταινίες κάποιου ηθοποιού συναντώνται πιθανώς και στα δύο σετ, δεν έχει νόημα ο αποκλεισμός του από κανένα από αυτά. Άλλωστε, η πραγματική βαθμολογία που θα έδινε ο χρήστης για κάποιον ηθοποιό/σκηνοθέτη δεν είναι γνωστή, άρα πράγματι ζητάμε από το SVM να εξάγει συμπεράσματα για το σύνολο των εμφανιζόμενων ηθοποιών/σκηνοθετών, σε οποιοδήποτε σετ και αν έχουν συμπεριληφθεί ταινίες του.

Στα διαγράμματα που ακολουθούν καθεμία από τις 3 καμπύλες ενός διαγράμματος αφορά μια διαφορετική ταξινόμηση αυτών των συντελεστών.

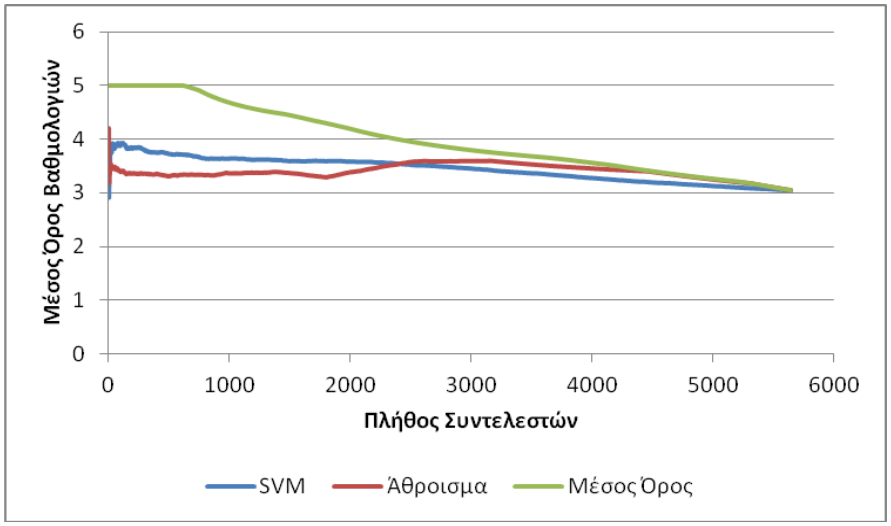
Η πράσινη γραμμή (Μέσος Όρος) αντιστοιχεί στην ταξινόμηση των συντελεστών με κριτήριο το μέσο όρο των βαθμολογιών των ταινιών τους.

Η κόκκινη γραμμή (Άθροισμα) αντιστοιχεί στην ταξινόμηση των συντελεστών με κριτήριο το άθροισμα των βαθμολογιών των ταινιών τους.

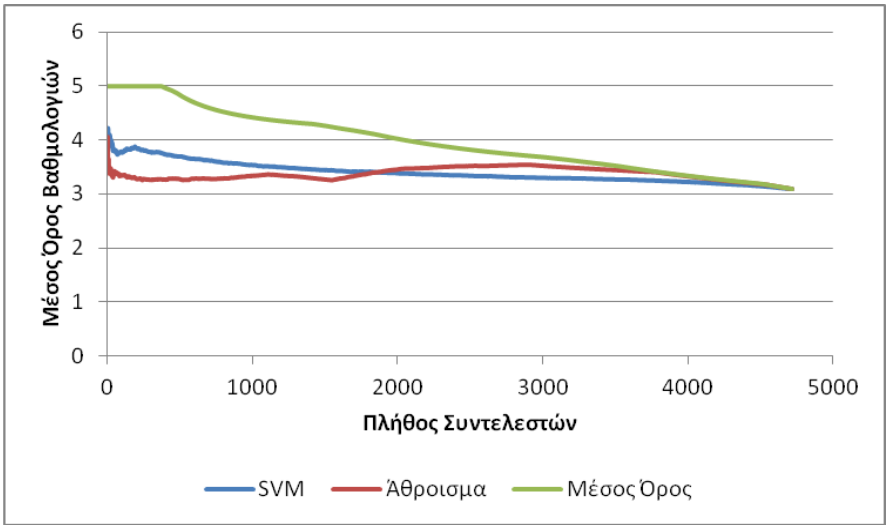
Τέλος, η μπλε γραμμή (SVM) αντιστοιχεί στην ταξινόμηση των συντελεστών από το Ranking SVM. Το σημείο (x,y) μιας καμπύλης αντιστοιχεί στο μέσο όρο y των συντελεστών που βρίσκονται στις πρώτες x θέσεις της λίστας συντελεστών όπως αυτή έχει διαμορφωθεί για καθεμία από τις διαφορετικές ταξινομήσεις.

Παρατίθενται στη συνέχεια διαγράμματα για επιλεγμένους χρήστες, παρόλο που αντίστοιχα πειράματα έχουν πραγματοποιηθεί και για τους 25 χρήστες που είδαμε στην προηγούμενη ενότητα. Ο λόγος για αυτή την επιλογή είναι η παρόμοια απόδοση της ταξινόμησης του SVM για όλους τους χρήστες, οπότε δεν κρίθηκε σκόπιμη η παράθεση όλων των διαγραμμάτων.

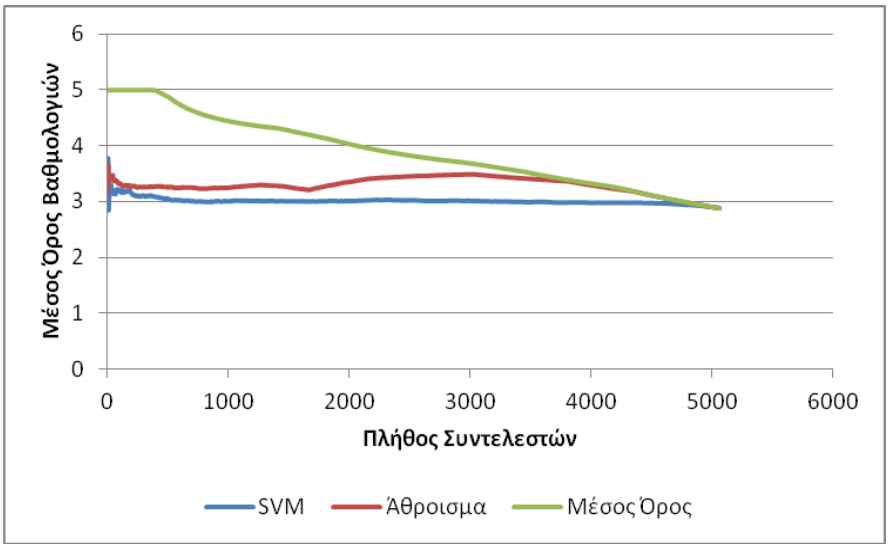
#### *7.2.2.1 Επιλεγμένα διαγράμματα χρηστών με πολύ μεγάλο πλήθος βαθμολογιών*



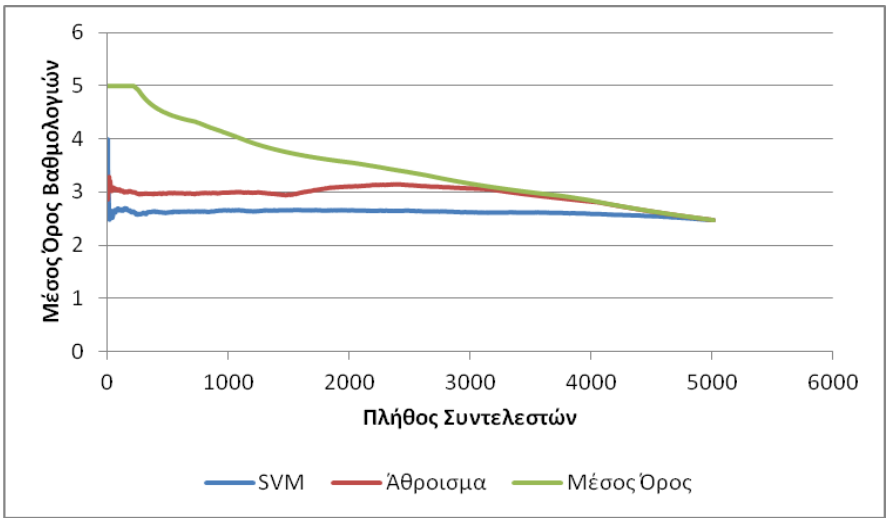

**Διάγραμμα 29: rerank συντελεστών ταινιών για το χρήστη A8**



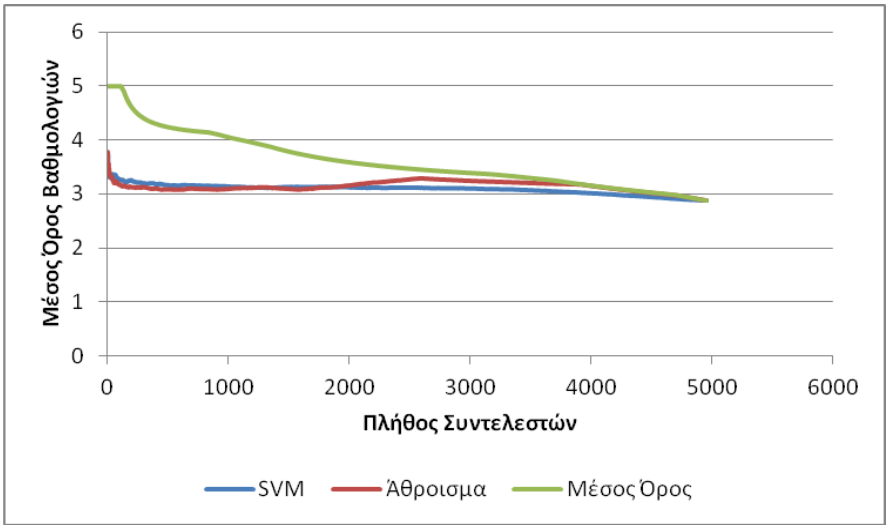

**Διάγραμμα 30: rerank συντελεστών ταινιών για το χρήστη Α7**



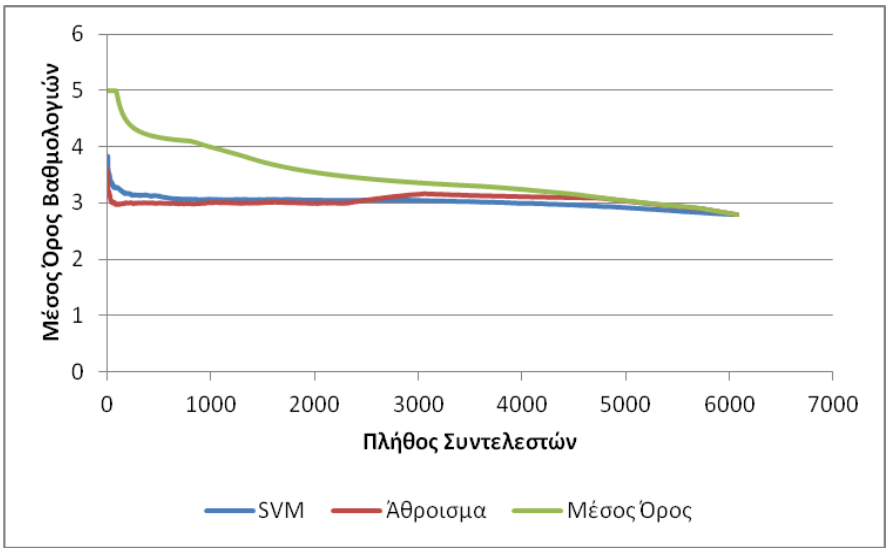

**Διάγραμμα 31: rerank συντελεστών ταινιών για το χρήστη Α6**




**Διάγραμμα 32: rerank συντελεστών ταινιών για το χρήστη Α4**



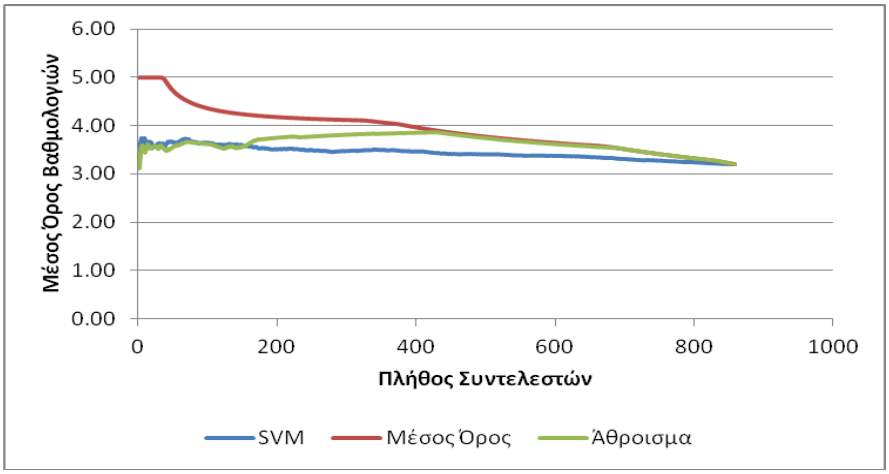

**Διάγραμμα 33: rerank συντελεστών ταινιών για το χρήστη Α3**



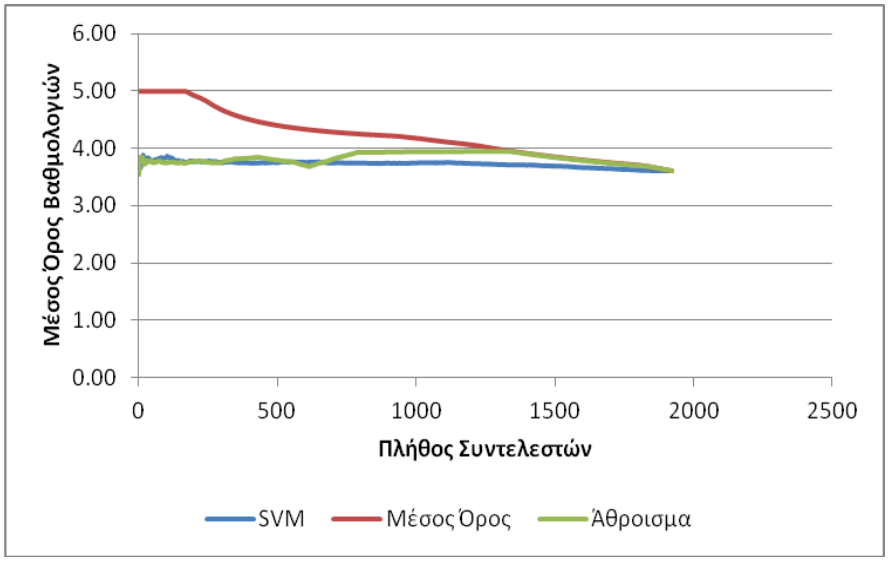

**Διάγραμμα 34:rerank συντελεστών ταινιών για το χρήστη Α1**

Εξαιτίας της παρόμοιας συμπεριφοράς της μεθόδου σε αυτό το σύνολο χρηστών με το σύνολο των χρηστών με λιγότερες βαθμολογίες, ο σχολιασμός θα γίνει στην επόμενη ενότητα για όλα τα διαγράμματα μαζί.

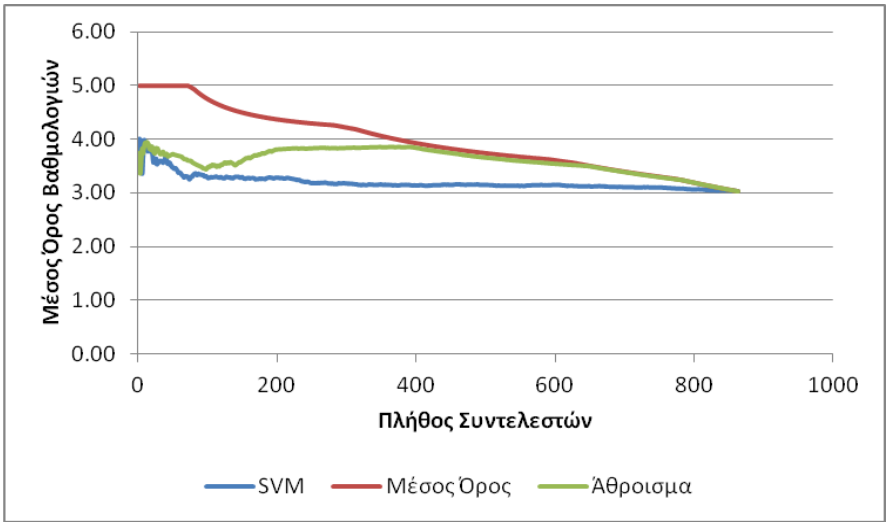
### 7.2.2.2 *Επιλεγμένα διαγράμματα χρηστών με σχετικά μικρό πλήθος βαθμολογιών*



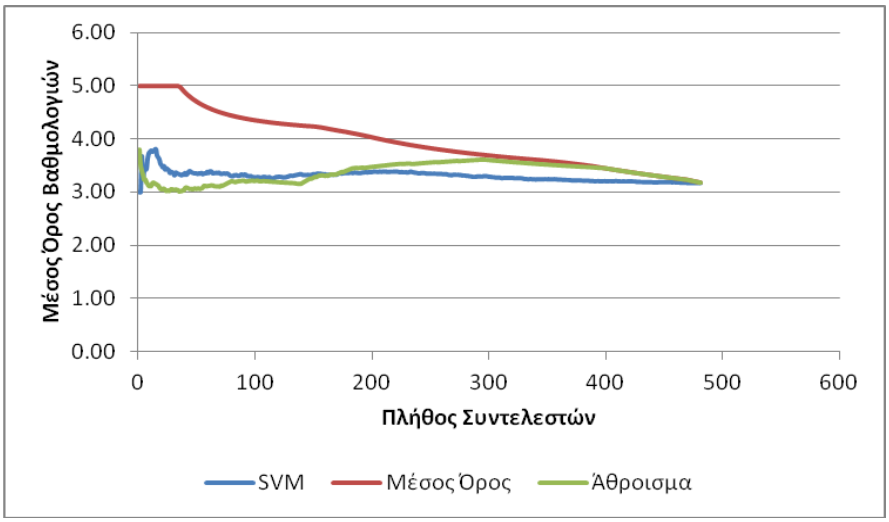

**Διάγραμμα 35: rerank συντελεστών ταινιών για το χρήστη B2**




**Διάγραμμα 36: rerank συντελεστών ταινιών για το χρήστη B10**




**Διάγραμμα 37: rerank συντελεστών ταινιών για το χρήστη B7**




**Διάγραμμα 38: rerank συντελεστών ταινιών για το χρήστη B8**

Παρατηρήσεις (αφορούν και τα δύο σύνολα χρηστών)

- Βλέπουμε πως ακόμα και στους χρήστες με μικρό σχετικά πλήθος βαθμολογημένων ταινιών, το σύνολο των ηθοποιών και των σκηνοθετών που εμφανίζονται σε αυτές είναι πολύ μεγάλο. Δεδομένου ότι η εκπαίδευση του SVM για την αναταξινόμηση τέτοιων αποτελεσμάτων γίνεται με έμμεσο τρόπο, καταλαβαίνουμε πως η συγκρότηση ενός τόσο μεγάλου συνόλου δεδομένων για αξιολόγηση είναι πολύ πιθανό να επιδρά αρνητικά στην απόδοση της μεθόδου
- Πράγματι, η μπλε καμπύλη, η οποία αντιστοιχεί στην ταξινόμηση με χρήση του SVM, κινείται σε κάθε περίπτωση πολύ κοντά στο συνολικό μέσο όρο των βαθμολογιών των συντελεστών. Υπενθυμίζουμε ότι ως βαθμολογία ενός ηθοποιού/σκηνοθέτη έχουμε ορίσει το μέσο όρο των βαθμολογιών των ταινιών στις οποίες συμμετείχε.
- Σε αντίθεση με τα διαγράμματα για την ταξινόμηση των ταινιών, εδώ δεν εμφανίζεται η αντίστοιχη καμπύλη της ταξινόμησης των αποτελεσμάτων από τη μηχανή αναζήτησης. Ο λόγος για αυτό είναι η συμπεριφορά της καμπύλης του SVM από την οποία φαίνεται πως δεν υπάρχει ουσιαστική επίδραση της μεθόδου στην αναταξινόμηση.
- Βλέπουμε πως σε κάποια διαγράμματα η μπλε καμπύλη (SVM) κινείται πάνω ή πολύ κοντά στην πράσινη γραμμή (ταξινόμηση με χρήση αθροίσματος βαθμολογιών) ή και εναλλάσσονται. Η κόκκινη γραμμή (ταξινόμηση με χρήση μέσου όρου βαθμολογιών), όπως είναι φυσικό παραμένει πολύ ψηλότερα σε κάθε περίπτωση.

Θα εξετάσουμε τώρα τους πιθανούς παράγοντες που επηρέασαν την απόδοση της εξατομικευμένης ταξινόμησης ηθοποιών και σκηνοθετών.

- **Καταλληλότητα αρχικών δεδομένων**
  - Τα αρχικά μας δεδομένα δεν περιλαμβάνουν προτιμήσεις χρηστών για συντελεστές ταινιών, παρά μόνο βαθμολογίες για ταινίες. Συνεπώς η διαθέσιμη πληροφορία για τους ηθοποιούς και τους σκηνοθέτες εξήχθη με έμμεσο τρόπο από το SVM, γεγονός που εισάγει ένα σημαντικό παράγοντα σφάλματος. Η εκπαίδευση του SVM πραγματοποιήθηκε με ένα μόνο είδος δεδομένων, ενώ η αξιολόγησή της γίνεται και μέσω άλλου είδους δεδομένων. Τα χαρακτηριστικά που δημιουργήθηκαν για την εκπαίδευση προσαρμόστηκαν σε αυτή την επιθυμητή λειτουργία, ωστόσο πρέπει να αξιολογηθεί περαιτέρω αν μια τέτοια προσέγγιση μπορεί να εφαρμοστεί από το Ranking SVM.
  - Ένας άλλος σημαντικός παράγοντας είναι η απώλεια κάποιων ταινιών από τις αρχικά διαθέσιμες για κάθε χρήστη κατά τη διαδικασία αντιστοίχισης των δεδομένων του Netflix με εγγραφές της DBpedia. Οι χρήστες επιλέχθηκαν ώστε να έχει διατηρηθεί μεγάλο ποσοστό των βαθμολογιών τους, ωστόσο είναι πιθανό οι βαθμολογίες των ταινιών που χάθηκαν να έχουν επηρεάσει τη συνοχή του προφίλ κάποιου χρήστη. Αυτός είναι ένας παράγοντας που δεν υπήρχε τρόπος να ρυθμιστεί κατά την επιλογή χρηστών.
  - Τέλος, στις πληροφορίες της DBpedia σχετικά με τους ηθοποιούς μιας ταινίας δε γίνεται διαχωρισμός των βασικών συντελεστών μιας ταινίας. Τα infoboxes των άρθρων της Wikipedia από όπου εξάγονται οι πληροφορίες της dbpedia περιλαμβάνουν τους ηθοποιούς που συμμετείχαν σε μια ταινία, χωρίς να διαχωρίζουν τους πρωταγωνιστές και χωρίς να εξασφαλίζεται απαραίτητα ότι τα όνοματα που αναφέρονται περιλαμβάνουν τους βασικούς ηθοποιούς της ταινίας. Άλλωστε, όπως έχουμε ήδη αναφέρει η αξιοπιστία των δεδομένων της Wikipedia πρέπει να κρίνεται μέσα από το πρίσμα της ελεύθερης και ανοιχτής συνεργατικής δημιουργίας γνώσης. Είναι λογική η υπόθεση ότι οι πρωταγωνιστές μιας ταινίας παίζουν σημαντικότερο ρόλο στην αξιολόγησή της από κάποιο χρήστη. Ωστόσο, από όσα αναφέραμε γίνεται κατανοητό ότι ένας ηθοποιός που κατείχε δεύτερους ρόλους σε καλές, κατά το χρήστη, ταινίες θα θεωρηθεί ιδιαίτερα σημαντικός. Γίνεται αντιληπτό ότι αυτή δεν είναι πάντα η επιθυμητή κατάσταση και μπορεί να δώσει στο SVM παραπλανητική πληροφορία.

- **Εύρεση κατάλληλου μέτρου αξιολόγησης της μεθόδου**

  Ενώ για τις ταινίες διαθέτουμε ένα testing set που μπορεί να χρησιμοποιηθεί για την αξιολόγηση της ορθότητας της εξατομικευμένης ταξινόμησής τους από το SVM, δεν ισχύει το ίδιο για τους συντελεστές μιας ταινίας. Εκεί η κρίση μας βασίζεται ουσιαστικά

στο propagation της βαθμολογίας της ταινίας στους συντελεστές της. Στις γραφικές χρησιμοποιήθηκαν δύο μέτρα σύγκρισης: ο μέσος όρος των βαθμολογιών των ταινιών στις οποίες έχει συμμετάσχει ένας ηθοποιός/σκηνοθέτης και το άθροισμα των βαθμολογιών αυτών. Η πρώτη επιλογή είναι προφανής, η δεύτερη χρησιμοποιείται για να αποτυπώσει το συνολικό ενδιαφέρον που έχει δείξει ένας χρήστης για κάποιον ηθοποιό/σκηνοθέτη. Αν αναλογιστούμε το εξής σενάριο που αντιστοιχεί σε πραγματικό χρήστη και συγκεκριμένα στο χρήστη B4.

| ηθοποιός | πλήθος ταινιών | Μ.Ο. βαθμολογιών | άθροισμα βαθμολογιών |
|---|---|---|---|
| Clint Eastwood | 22 | 3.55 | 78 |
| Harvey Keitel | 1 | 5 | 5 |
| Kurt Russell | 10 | 3.3 | 33 |
| Brian Cox | 8 | 4 | 32 |

Είναι προφανές ότι κανένα από τα δύο μέτρα σύγκρισης δεν μπορεί απόλυτα να αποτυπώσει τις προτιμήσεις του χρήστη. Διαισθητικά, οι βαθμολογίες του χρήστη για 22 ταινίες του Clint Eastwood οι οποίες συγκεντρώνουν έναν ικανοποιητικό μέσο όρο (3.55)μπορούν να θεωρηθούν ένδειξη προτίμησης σε αυτόν. Ωστόσο, με κριτήριο το μέσο όρο ο Harvey Keitel θα έπρεπε να εμφανιστεί ψηλότερα στη λίστα αποτελεσμάτων. Αντίστοιχα, για τους Kurt Russell και Brian Cox το σχεδόν ίδιο άθροισμα βαθμολογιών μεταφράζεται σε πολύ διαφορετικό μέσο όρο άρα επίσης από μόνο του δεν μπορεί να αποτελεί κριτήριο.

Θα μπορούσε πιθανώς να χρησιμοποιηθεί ένα συνδυαστικό μέτρο αξιολόγησης. Για παράδειγμα για τον ίδιο χρήστη (B4) βρίσκουμε ότι υπάρχουν στις ταινίες του 1921 διαφορετικοί ηθοποιοί και σκηνοθέτες. Από αυτούς, 32 εμφανίζονται σε περισσότερες από 6 ταινίες του χρήστη και συγκεντρώνουν μέσο όρο μεγαλύτερο του 3.7. Με την ταξινόμηση του SVM, οι 8 από αυτούς τοποθετούνται στις πρώτες 15 θέσεις της λίστας αποτελεσμάτων (συγκεκριμένα στις θέσεις 1,2,5,10,11,12,13 και 15). Ωστόσο δεν είναι προφανές ποιες επιλογές πλήθος ταινιών και μέσου όρου πρέπει να γίνουν σε κάθε περίπτωση ώστε να έχουμε μια αντιπροσωπευτική εικόνα της απόδοσης της ταξινόμησης.

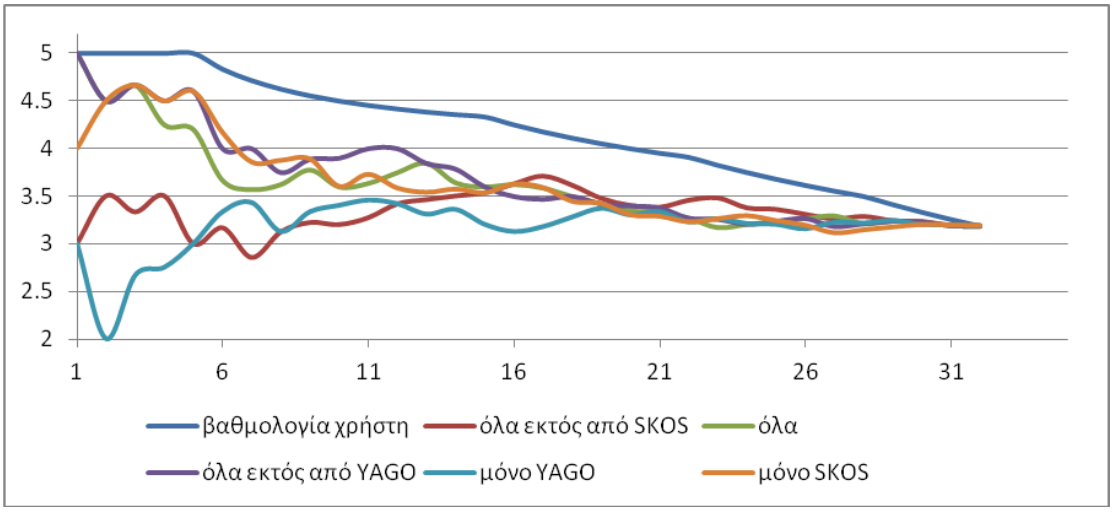
Ακόμα και με την εφαρμογή όμως ενός εναλλακτικού μέτρου αξιολόγησης της μεθόδου, δεν μπορεί να ξεπεραστεί το πρόβλημα που αναφέρθηκε παραπάνω σχετικά με την έλλειψη πληροφοριών για το διαχωρισμό των συντελεστών μιας ταινίας σε λιγότερο και περισσότερο σημαντικούς. Ο παράγοντας αυτός επηρεάζει τόσο την αξιοπιστία των πληροφοριών με τις οποίες εκπαιδεύεται το SVM, αλλά παράλληλα και το μέτρο αξιολόγησης που χρησιμοποιείται.

### *7.2.3 Εκπαίδευση Ranking SVM με διαφορετικές ομάδες χαρακτηριστικών*

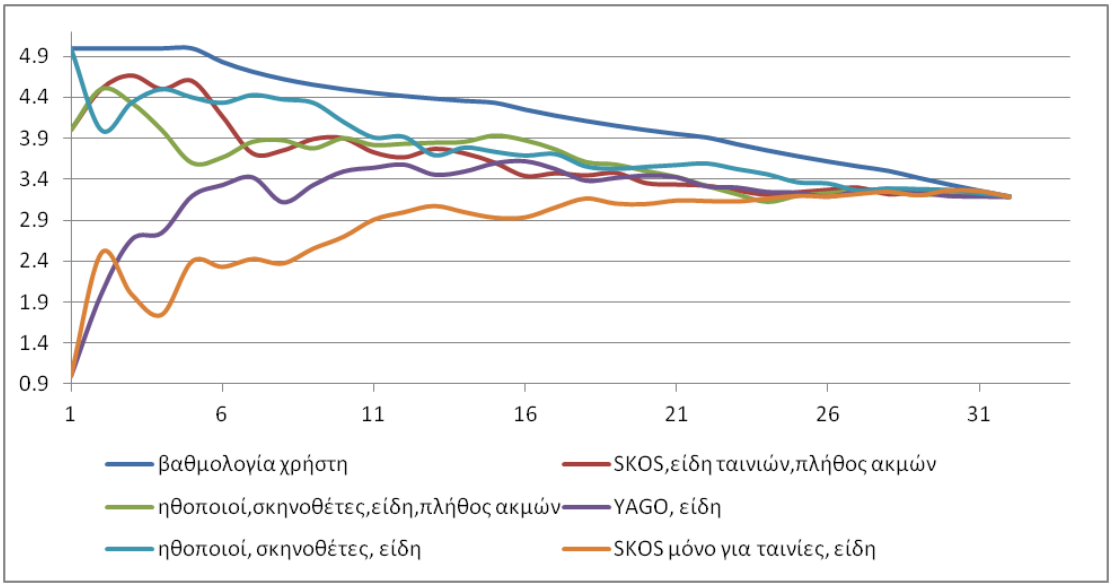
Περιγράψαμε σε προηγούμενη ενότητα τις 10 διαφορετικές ομάδες χαρακτηριστικών που δημιουργήθηκαν για την εκπαίδευση του SVM. Στην προηγούμενη παράγραφο είδαμε την απόδοση της μεθόδου όταν στην εκπαίδευση χρησιμοποιείται το σύνολο των χαρακτηριστικών, δηλαδή και οι 10 αυτές ομάδες. Πραγματοποιήσαμε επιπλέον πειράματα με τη χρήση διαφόρων συνδυασμών αυτών των ομάδων.

Παραθέτουμε ενδεικτικά, για το χρήστη B3, διαγράμματα που δείχνουν την απόδοση της μεθόδου για 10 διαφορετικούς συνδυασμούς αυτών των ομάδων.

Σε αντίθεση με τα διαγράμματα της προηγούμενης παραγράφου δε συμπεριλαμβάνεται εδώ η καμπύλη που αντιστοιχεί στην ταξινόμηση των αποτελεσμάτων από τη μηχανή αναζήτησης, αφού σκοπός μας είναι να φανούν οι διαφορές από την προσθήκη ή αφαίρεση ομάδων χαρακτηριστικών κατά την εκπαίδευση. Για λόγους πληρότητας, όμως, έχει συμπεριληφθεί και στα δύο διαγράμματα η καμπύλη που εκφράζει τις πραγματικές προτιμήσεις του χρήστη.




**Διάγραμμα 39: εκπαίδευση με διαφορετικές ομάδες χαρακτηριστικών (α)**




**Διάγραμμα 40: εκπαίδευση με διαφορετικές ομάδες χαρακτηριστικών (β)**

Τα καλύτερα αποτελέσματα για το συγκεκριμένο χρήστη φαίνεται να προκύπτουν από τη χρήση όλων των ομάδων εκτός από τις δύο ομάδες που σχετίζονται με τις κλάσεις της οντολογίας SKOS. Βλέπουμε μάλιστα πως για το συγκεκριμένο χρήστη τα χαρακτηριστικά των ομάδων αυτών γενικά επηρεάζουν αρνητικά την εκπαίδευση του Ranking SVM.

Ωστόσο πρέπει να παρατηρήσουμε πως ενώ η εκπαίδευση με χαρακτηριστικά (σχεδόν) αποκλειστικά των 2 αυτών ομάδων έχει ιδιαίτερα αρνητικά αποτελέσματα (καμπύλη "μόνο YAGO" στο διάγραμμα 36 , "YAGO, είδη" στο 37), η επίδρασή τους σχεδόν χάνεται όταν για την εκπαίδευση χρησιμοποιηθούν τα χαρακτηριστικά όλων των ομάδων. Εκεί βλέπουμε πως τελικά υπερισχύουν τα χαρακτηριστικά που δίνουν μια αντιπροσωπευτική των προτιμήσεων του χρήστη ταξινόμηση.

Αυτή η παρατήρηση είναι σταθερή για όλους τους χρήστες που χρησιμοποιήθηκαν στα πειράματα. Ενώ το καλύτερο υποσύνολο ομάδων χαρακτηριστικών μπορεί να διαφέρει σημαντικά για τους διάφορους χρήστες, η χρήση όλων των ομάδων είχε πάντα σταθερά καλά αποτελέσματα, ήταν δηλαδή πάντα η καλύτερη ή πολύ κοντά στην καλύτερη επιλογή.

### *7.2.4 Παράδειγμα χρήσης του Ranking SVM σε αναζήτηση*

Ερώτημα: Christopher Walken , film

Για το ερώτημα αυτό ορίσαμε μέγιστο επιθυμητό μήκος μονοπατιού ίσο με 3, δηλαδή θέλουμε να βρούμε μονοπάτια με μια μόνο ακμή. Σκοπός του ερωτήματος είναι να

ανακτηθούν οι ταινίες στις οποίες έχει παίξει ο Christopher Walken. Για το ερώτημα αυτό η μηχανή αναζήτησης επιστρέφει περισσότερα από 50 μονοπάτια. Είναι προφανές πως αν αυξήσουμε το μήκος μονοπατιού ο αριθμός αυτός θα αυξηθεί σημαντικά. Επιπλέον, αν προσθέσουμε μια ακόμα φράση κλειδί, όπως για παράδειγμα το όνομα ενός ηθοποιού, είναι πιθανό να αυξήσουμε τόσο το πλήθος των αποτελεσμάτων όσο και το μέγεθος κάθε ξεχωριστού αποτελέσματος, στην περίπτωση που βρεθούν συνδυασμοί μονοπατιών που να περιλαμβάνουν όλους τους κόμβους που αντιστοιχούν στις φράσεις-κλειδιά.

Ένα υποσύνολο των αποτελεσμάτων αποτελούν τα εξής 4 μονοπάτια, τα οποία εμφανίζονται στα αποτελέσματα με τη σειρά που δίνεται εδώ:

<Christopher Walken> <starring> <The Stepford Wives>

<Christopher Walken> <starring> <Around The Bend>

<Christopher Walken> <starring> <Man On Fire>

<Christopher Walken> <starring> <Communion>

Οι βαθμολογίες του χρήστη A4 για τις παραπάνω ταινίες είναι:

| | |
|---|---|
| <The Stepford Wives> | 1 |
| <Around The Bend> | 4 |
| <Man On Fire> | 5 |
| <Communion> | 2 |

Οι ταινίες αυτές περιλαμβάνονται στο testing set του χρήστη A4, οπότε έχει ενδιαφέρον να δούμε σε ποια σειρά τοποθετούνται αυτές από το SVM. Χρησιμοποιούμε το SVM που έχει εκπαιδευτεί με όλα τα χαρακτηριστικά που έχουν ορισθεί. Μετά την εφαρμογή της αναταξινόμησης , η σειρά εμφάνισης των μονοπατιών είναι η εξής:

<Christopher Walken> <starring> <Man On Fire>

<Christopher Walken> <starring> <Around The Bend>

<Christopher Walken> <starring> <Communion>

<Christopher Walken> <starring> <The Stepford Wives>

η οποία όπως βλέπουμε ανταποκρίνεται απόλυτα στις πραγματικές βαθμολογίες του χρήστη.

Στο παράδειγμα αυτό καθένα από τα μονοπάτια διαφοροποιείται σε έναν μόνο κόμβο και όλοι αυτοί οι κόμβοι έχουν λάβει το ίδιο αρχικό σκορ (lucene score και αναζήτηση στα abstracts). Έτσι, η εξατομικευμένη ταξινόμησή τους ισοδυναμεί με την ταξινόμηση των διαφορετικών αυτών κόμβων.

Θα πρέπει να τονίσουμε ότι εδώ μας ενδιαφέρουν αποτελέσματα που να περιλαμβάνονται στο testing set του χρήστη ώστε να μπορούμε να αξιολογήσουμε τη συμπεριφορά της μεθόδου. Αποδείχθηκε μάλιστα ιδιαίτερα επίπονη η διαδικασία εύρεσης κατάλληλων ερωτημάτων ώστε η εξατομικευμένη ταξινόμησή τους να μπορεί να αξιολογηθεί από τα δεδομένα του testing set. Σε πραγματικές συνθήκες εφαρμογής της μεθόδου, το σύστημα θα επαναταξινομήσει το σύνολο των 50 μονοπατιών, οπότε αντιλαμβανόμαστε καλύτερα την αξία της εξατομικευμένης αυτής ταξινόμησης ώστε ο χρήστης να βρει στις πρώτες θέσεις της λίστας τα αποτελέσματα που είναι πιο κοντά στα ενδιαφέροντά του.

## *7.3 Σύνοψη συμπερασμάτων αξιολόγησης*

Από τα πειράματα που πραγματοποιήθηκαν μπορούν να εξαχθούν συνοπτικά τα εξής συμπεράσματα:

- Στην εξατομικευμένη ταξινόμηση ταινιών η χρήση του Ranking SVM με τα χαρακτηριστικά που δημιουργήσαμε μπορεί πράγματι στην πλειοψηφία των περιπτώσεων να φέρει πιο ψηλά στη λίστα αποτελεσμάτων τις ταινίες που ανταποκρίνονται περισσότερο στις προτιμήσεις του χρήστη, βελτιώνοντας σημαντικά την ταξινόμηση της μηχανής αναζήτησης.
- Η απόδοση της μεθόδου παρουσιάζει σημαντική βελτίωση για χρήστες με μεγαλύτερο σύνολο δεδομένων εκπαίδευσης.
- Στην εξατομικευμένη ταξινόμηση συντελεστών ταινιών παρουσιάζονται προβλήματα λόγω της αρχικής έλλειψης αντίστοιχων δεδομένων εκπαίδευσης, αλλά και της δυσκολίας στην εύρεση ενός αξιόπιστου μέτρου αξιολόγησης της μεθόδου
- Η χρήση διαφορετικών ομάδων χαρακτηριστικών μπορεί να επηρεάσει την απόδοση της μεθόδου. Το μέγεθος της επίδρασης μπορεί να διαφέρει σημαντικά ανάλογα το χρήστη και το συνδυασμό χαρακτηριστικών που εξετάζεται κάθε φορά. Η χρήση όλων των χαρακτηριστικών έχει συνήθως τα καλύτερα αποτελέσματα ή ελάχιστα χειρότερα από αυτά του καλύτερου συνδυασμού, οπότε αποτελεί μια ασφαλή επιλογή για την εκπαίδευση του SVM.
- Για τους χρήστες που η μέθοδος δεν έχει ουσιαστικά αποτελέσματα με τη χρήση όλων των χαρακτηριστικών, δε βρέθηκε κάποιο υποσύνολό τους που να επηρεάζει θετικά την απόδοσή της. Εδώ πρέπει να εξεταστούν πιο αναλυτικά οι πιθανές ιδιαιτερότητες του συνόλου δεδομένων αυτών των χρηστών. Πρέπει ωστόσο να ληφθεί υπόψη η απώλεια δεδομένων κατά την αντιστοίχιση των δεδομένων του Netflix Prize σε εγγραφές της

DBpedia, καθώς είναι πιθανό οι ταινίες που δεν ανακτήθηκαν να επηρέασαν σημαντικά τη συνοχή του προφίλ κάποιων χρηστών.

# 8

# *Επίλογος*

Σε αυτό το κεφάλαιο γίνεται μια σύντομη ανασκόπηση των συμπερασμάτων που εξήχθησαν στα πλαίσια της παρούσας διπλωματικής και προτείνονται πιθανές επεκτάσεις της με βάση προβληματισμούς που προέκυψαν κατά την εκπόνησή της.

## *8.1 Σύνοψη και συμπεράσματα*

Αντικείμενο της παρούσας διπλωματικής ήταν η προσαρμογή και αξιολόγηση μεθόδων εξατομίκευσης αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα.

Στο πρώτο στάδιο μελετήθηκε η χρήση μιας μηχανής αναζήτησης σε σημασιολογικά δεδομένα, επεκτάθηκε η λειτουργία της και καταγράφηκαν τα προβλήματα που εισάγονται από το rdf μοντέλο στην αναζήτηση με λέξεις κλειδιά.

Στη συνέχεια, εμπλουτίστηκε με σημασιολογικές πληροφορίες ένα σύνολο δεδομένων που περιλάμβανε αποκλειστικά ταινίες και βαθμολογίες μαζί με τις ημερομηνίες που αυτές δόθηκαν από τους χρήστες. Το νέο αυτό σύνολο δεδομένων χρησιμοποιήθηκε για τη δημιουργία προφίλ χρηστών στα οποία να αποτυπώνονται τα ενδιαφέροντά τους.

Δημιουργήθηκε ένας μεγάλος αριθμός χαρακτηριστικών εκπαίδευσης του Ranking SVM, δηλαδή του συστήματος εξατομίκευσης αναζήτησης και αξιολογήθηκε η

αποτελεσματικότητά τους στην εξατομικευμένη ταξινόμηση αποτελεσμάτων αναζήτησης που περιλαμβάνουν ταινίες, ηθοποιούς και σκηνοθέτες.

Τα πειράματα έδειξαν ότι ειδικά στα αποτελέσματα που αφορούν ταινίες, η μέθοδος βελτιώνει σημαντικά τη σειρά επιστροφής των αποτελεσμάτων της αναζήτησης ώστε να είναι αντιπροσωπευτική των προτιμήσεων του χρήστη.

## *8.2 Προβληματισμοί και Μελλοντικές Επεκτάσεις*

### *8.2.1 Καταλληλότητα δεδομένων εισόδου*

Είδαμε σε προηγούμενο κεφάλαιο πως εμπλουτίστηκε το σύνολο των αρχικών δεδομένων με σημασιολογικές πληροφορίες με στόχο το σχηματισμό προφίλ χρηστών που θα καταδεικνύουν περισσότερα στοιχεία για τις προτιμήσεις τους. Παραμένει ωστόσο το ερώτημα αν τελικά από απλές βαθμολογίες ταινιών μπορούν να εξαχθούν αξιόπιστα συμπεράσματα για προτιμήσεις σε ηθοποιούς, σκηνοθέτες και γενικότερα για ζητήματα που κινούνται μεν στο χώρο των ταινιών, αλλά σε ένα ευρύτερο πλαίσιο.

Γενικότερα, μπορεί να μελετηθεί η δυνατότητα επέκτασης δεδομένων που προορίζονται για αλγόριθμους συνεργατικού φιλτραρίσματος σε content-based συστήματα εξατομίκευσης αναζήτησης μέσω του εμπλουτισμού τους με σημασιολογικές πληροφορίες.

### *8.2.2 Προσαρμογή SVM*

Τα δεδομένα εκπαίδευσης του SVM περιλαμβάνουν μόνο βαθμολογίες ταινιών. Σκοπός μας είναι μετά την εκπαίδευσή του το SVM να χρησιμοποιηθεί για την εξατομικευμένη αναταξινόμηση και άλλων οντοτήτων και συγκεκριμένα στα δικά μας πειράματα ηθοποιών και σκηνοθετών. Με αυτό το κριτήριο δημιουργήθηκαν χαρακτηριστικά εκπαίδευσης που να έχει νόημα να εφαρμοσθούν και για τις τρεις αυτές κατηγορίες οντοτήτων. Σημειώνουμε εδώ ότι σκοπός μας είναι το SVM να χρησιμοποιηθεί για την αναταξινόμηση λίστας αποτελεσμάτων που περιλαμβάνει ένα μόνο είδος των παραπάνω τριών κατηγοριών κάθε φορά, δηλαδή μια λίστα ηθοποιών/σκηνοθετών/ταινιών ξεχωριστά. Είναι προφανές ότι δεν έχουμε τις κατάλληλες πληροφορίες ώστε να συγκρίνουμε το ενδιαφέρον ενός χρήστη για μια ταινία σε σχέση με το ενδιαφέρον του για έναν ηθοποιό. Ωστόσο παραμένει ανοιχτό το ερώτημα αν πράγματι μπορούν να βρεθούν τα κατάλληλα χαρακτηριστικά για έναν τέτοιο

σκοπό ή αν το SVM μπορεί να αναταξινομεί επιτυχώς μόνο δεδομένα αντίστοιχα με αυτά για τα οποία εκπαιδεύτηκε χωρίς δυνατότητες για μια τέτοια επέκταση.

Υπενθυμίζουμε ότι για την προσέγγισή μας ακολουθήθηκε μια ιδιαίτερη μέθοδος ανάθεσης τιμών στα χαρακτηριστικά για τα αποτελέσματα διαφορετικού τύπου.

Ιδιαίτερα μπορεί να μελετηθεί η σημασία της μηδενικής τιμής ενός χαρακτηριστικού για κάποιο αποτέλεσμα της λίστας αποτελεσμάτων. Η μηδενική τιμή μπορεί να προκύψει γιατί πραγματικά ένα χαρακτηριστικό παίρνει την τιμή 0 ή γιατί η τιμή του είναι άγνωστη. Στο δικό μας τρόπο ανάθεσης τιμών η μηδενική τιμή μπορεί να εκφράζει και πως ένα χαρακτηριστικό είναι δομικά απόν (δηλαδή δεν εφαρμόζεται) σε κάποιο αποτέλεσμα, όπως για παράδειγμα η μηδενική τιμή ενός ηθοποιού σε χαρακτηριστικό που αντιστοιχεί στο όνομα ενός άλλου ηθοποιού.

### *8.2.3 Παράγοντες που επηρεάζουν την απόδοση του SVM στην αναταξινόμηση ταινιών*

Είδαμε πως στη γενική περίπτωση η εφαρμογή του Ranking SVM βελτιώνει το μέσο όρο των αποτελεσμάτων που τοποθετούνται ψηλότερα στη λίστα αποτελεσμάτων σε σύγκριση με την αντίστοιχη λίστα αποτελεσμάτων της μηχανής αναζήτησης. Ωστόσο υπάρχουν χρήστες στους οποίους η χρήση εξατομίκευσης δεν έχει ορατά αποτελέσματα, χωρίς να γνωρίζουμε για ποιο λόγο συμβαίνει αυτό.

Παρόλο που τα πειράματα έδειξαν ότι ένα μεγαλύτερο training set μπορεί να εξασφαλίσει καλύτερα αποτελέσματα και μεγαλύτερη σταθερότητα στην απόδοση του SVM, η ύπαρξη χρηστών με αρκετά πλούσιο σύνολο βαθμολογιών για τους οποίους η μέθοδος δεν είχε ορατά αποτελέσματα δείχνει πως ένα μεγάλο σύνολο δεδομένων δεν εξασφαλίζει απαραίτητα την ορθή αναταξινόμηση. Προτείνουμε κάποιους πιθανούς παράγοντες που επηρεάζουν την απόδοση του SVM.

- **Εξάρτηση από το σύνολο δεδομένων κάποιου χρήστη**

Η απώλεια δεδομένων κατά την αντιστοίχιση των ταινιών του Netflix σε εγγραφές της DBpedia είναι ένας παράγοντας που πρέπει να ληφθεί υπόψη. Είναι πιθανό σε ορισμένες περιπτώσεις οι ταινίες που δεν ήταν εφικτό να αντιστοιχιστούν σε εγγραφές της DBpedia να ήταν σημαντικές για τη συνοχή του προφίλ ενός χρήστη και για το λόγο αυτό η απώλεια δεδομένων να επηρέασε σημαντικά την αποτελεσματικότητα της μεθόδου.

Επιπλέον, είναι πιθανό η εκπαίδευση του SVM να επηρεάζεται από την ποικιλομορφία των δεδομένων εκπαίδευσης ή να υπάρχει ασυνέπεια μεταξύ δεδομένων εκπαίδευσης και αξιολόγησης για κάποιο χρήστη.

- Χαρακτηριστικά SVM

Στα πλαίσια της παρούσας διπλωματικής δημιουργήθηκε ένας μεγάλος αριθμός χαρακτηριστικών για την εκπαίδευση του SVM στην αναταξινόμηση εγγραφών της DBpedia σχετικών με το χώρο του κινηματογράφου. Τα χαρακτηριστικά αυτά χωρίστηκαν σε 10 βασικές ομάδες. Η μέθοδος είχε ικανοποιητικά αποτελέσματα, κυρίως στην εξατομικευμένη ταξινόμηση ταινιών. Παρόλο που δοκιμάστηκε η χρήση διαφορετικών ομάδων χαρακτηριστικών και συνδυασμών τους στην εκπαίδευση του SVM, σε κάθε περίπτωση τα χαρακτηριστικά μιας ομάδας περιλαμβάνονταν είτε στο σύνολό τους είτε καθόλου.

Ως μελλοντική επέκταση, προτείνεται η εξέταση της επιλογής χαρακτηριστικών (feature selection [SOLW08], [TWT10]) και εντός της ίδιας ομάδας. Ειδικότερα, θα είχε ενδιαφέρον η ανάπτυξη μεθόδου πρόβλεψης των κατάλληλων χαρακτηριστικών εκπαίδευσης, με κριτήριο τις ιδιαιτερότητες του συνόλου δεδομένων κάθε χρήστη ξεχωριστά. Η σωστή επιλογή χαρακτηριστικών μπορεί να εξασφαλίσει τη σύγκλιση του αλγόριθμου εκπαίδευσης και την ορθότερη ταξινόμηση αποτελεσμάτων ώστε να είναι πιο αντιπροσωπευτική των προτιμήσεων του χρήστη.

Επιπλέον, μπορεί να μελετηθεί αν η καταλληλότερη επιλογή χαρακτηριστικών εκπαίδευσης για την εξατομικευμένη ταξινόμηση ταινιών συνεπάγεται ή όχι καλύτερα αποτελέσματα στην ταξινόμηση ηθοποιών και σκηνοθετών.

### *8.2.4 Αξιοποίηση της ημερομηνίας που δόθηκε η κάθε βαθμολογία*

Στα πειράματά μας το σύνολο των διαθέσιμων βαθμολογιών χωρίστηκε σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης με χρονικά κριτήρια, έτσι ώστε τα δεδομένα της εκπαίδευσης να προηγούνται χρονικά. Ωστόσο, στο [PC09] εξετάζεται η εξαγωγή επιπλέον συμπερασμάτων από την αξιοποίηση αυτών των πληροφοριών. Συγκεκριμένα, εξετάζονται οι πληροφορίες που προκύπτουν έμμεσα από τις ημερομηνίες που δόθηκαν στο σύστημα οι βαθμολογίες ενός χρήστη. Υπενθυμίζουμε ότι τα δεδομένα μας προέρχονται από το διαγωνισμό Netflix Prize. Στοιχεία όπως η χρονική απόσταση μιας βαθμολογίας ενός χρήστη από την πρώτη βαθμολογία που είχε δώσει στο σύστημα, το πλήθος των βαθμολογιών που δόθηκαν την ίδια μέρα, η συχνότητα με την οποία ο χρήστης εισάγει νέες βαθμολογίες στο σύστημα, μπορούν να αποκαλύψουν αν μια βαθμολογία αντιστοιχεί σε ταινία που είδε πρόσφατα ο χρήστης ή αν ο χρήστης δίνει στοιχεία στο σύστημα ώστε να βελτιωθούν οι ταινίες που του προτείνονται.

Η εξαγωγή τέτοιων πληροφοριών μπορεί να βελτιώσει τον τρόπο διαχωρισμού των ταινιών σε training και testing set ή και να αξιοποιηθεί στην εκπαίδευση του συστήματος

εξατομίκευσης ώστε να λαμβάνεται υπόψη η χρονική εξέλιξη των προτιμήσεών (πιθανώς με τη χρήση ενός short-term και ενός long-term ιστορικού).

### *8.2.5 Εύρεση αξιόπιστου τρόπου αξιολόγησης της εξατομικευμένης ταξινόμησης ηθοποιών και σκηνοθετών*

Στο κεφάλαιο 7 αναλύθηκαν οι δυσκολίες στην εύρεση ενός κατάλληλου τρόπου αξιολόγησης της εξατομίκευσης ταξινόμησης ηθοποιών και σκηνοθετών, αφού δεν είναι γνωστή η πραγματική γνώμη ενός χρήστη. Προτάθηκαν τρεις τρόποι και εξηγήθηκαν οι αδυναμίες καθενός σε συνδυασμό με τα προβλήματα που εισάγει η ελλιπής διαθέσιμη πληροφορία από τα αρχεία της DBpedia. Παραμένει ωστόσο ανοιχτό το ερώτημα αν υπάρχει μέτρο αξιολόγησης κατάλληλο για τα δεδομένα μας και κατά πόσο μπορούμε να εμπιστευτούμε τις βαθμολογίες των ταινιών ενός ηθοποιού/σκηνοθέτη ως ένδειξη της προτίμησης του χρήστη σε αυτόν.

# 9

# *Βιβλιογραφία*


[DEL+08] M. Dudev, S. Elbassuoni, J. Luxemburger, M. Ramanath and G. Weikum. Personalizing the Search for Knowledge. 2nd International Workshop on Personalized Access, Profile Management, and Context Awareness: Databases (PersDB'08), August 23, 2008, Auckland, New Zealand.

[MBH+09] E. Meij et al. Learning Semantic Query Suggestions. ISWC '09 Proceedings of the 8th International Semantic Web Conference, pp 424 - 440, Springer-Verlag Berlin, Heidelberg, 2009.

[SMB07] A. Sieg, B. Mobasher and R. Burke. Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search, in IEEE Intelligent Informatics Bulletin, Vol. 8, No. 1, November 2007.

[Sieg+07] A. Sieg, B. Mobasher and R. Burke. Web Search Personalization with Ontological User Profiles, in *CIKM'07,* November 6–8, 2007, Lisboa, Portugal.

[CLO11] F. Cena, S. Likavec and F. Osborne. Propagating user interests in ontology-based user model, in AI*IA'11 Proceedings of the 12th international conference on Artificial intelligence around man and beyond, pp 299-311, Springer-Verlag Berlin, Heidelberg, 2011.

[RSP04] C. Rocha, D. Schwabe, and M. P. Poggi. Hybrid approach for searching in the semantic web, in the WWW '04 Proceedings of the 13th international conference on World Wide Web, pp 374 - 383, ACM New York, NY, USA, 2004.

[DF11] L. Dali and B. Fortuna. Learning to Rank for Semantic Search, in the WWW'11 Proceedings, March 28th– April 1st, 2011, Hyderabad, India.

[DFTM12] L. Dali and B. Fortuna, T. Tran and D. Mladenić. Query-Independent learning to rank for RDF entity search, in ESWC'12 Proceedings of the 9th international conference on The Semantic Web: research and applications, pp 484-498, Springer-Verlag Berlin, Heidelberg, 2012.

[JT09] X. Jiang and A.H. Tan. Learning and inferencing in user ontology for personalized Semantic Web search, Information Sciences: an International Journal, Vol. 179, Issue 16, pp 2794-2808, Elsevier Science Inc. New York, NY, USA, July, 2009.

[LLNF12] K. W.-T. Leung, D. L. Lee, W. Ng and H. Y. Fung. A Framework for Personalizing Web Search with Concept-Based User Profiles, ACM Transactions on Internet Technology, Vol. 11, No. 4, Article 17, March 2012.

[Dud08] M. Dudev, Personalization of Search on Structured Data, Master's Thesis, Universit¨at des Saarlandes, FR Informatik, Max-Planck-Institut f¨ur Informatik, AG5, August 2008.

[SOLW08] K.-Q. Shen, C.-J. Ong, X.-P. Li and E. P. V. Wilder-Smith. Feature Selection via Sensitivity Analysis of SVM Probabilistic Outputs, Machine Learning, Volume 70, Issue 1, pp 1-20, Springer, January 2008.

[TWT10] M. Tan, L. Wang and I.W. Tsang. Learning Sparse SVM for Feature Selection on Very High Dimensional Datasets, in Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010.

[CP11] N. Chen and V. K. Prasanna. Learning to Rank Complex Semantic Relationships. Technical Report, University of Southern California, November 2011.

[SKW07] F.M. Suchanek, G. Kasneci and G. Weikum. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW '07 Proceedings of the 16th international conference on World Wide Web, pp 697-706, ACM New York, NY, USA, 2007.

[DLB09] M. Daoud, L.-T. Lechani and M. Boughanem. Towards a graph-based user profile modeling for a session-based personalized search. Knowledge and

Information Systems, Volume 21, Issue 3, pp 365-398, Springer, 2009

[LHL01] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web, Scientific American, May 17, 2001

[Joa06] T. Joachims. Training Linear SVMs in Linear Time, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2006

[PC09] Martin Piotte, Martin Chabbert. The Pragmatic Theory solution to the Netflix Grand Prize, Pragmatic Theory Inc., Canada, August 2009