Πολυτεχνείο Κρήτης
Τμήμα Ηλεκτρονικών Μηχανικών & Μηχανικών Υπολογιστών

Σύστημα Αυτόματης Δημιουργίας και Αναπροσαρμογής Προφίλ Χρηστών Σε περιβάλλον Ψηφιακής Τηλεόρασης

Ιωάννης-Αριστοτέλης Κοτόπουλος

Χανιά 2003

Σύστημα Αυτόματης Δημιουργίας και Αναπροσαρμογής Προφίλ Χρηστών Σε περιβάλλον Ψηφιακής Τηλεόρασης

Ιωάννης-Αριστοτέλης Κοτόπουλος

Μια εργασία που παρουσιάστηκε στο Πολυτεχνείο Κρήτης
για την εκπλήρωση των απαιτήσεων απόκτησης Διπλώματος στο
τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

Χανιά, Σεπτέμβρης 2003

Περίληψη

Η ψηφιακή τηλεόραση έρχεται για να βελτιώσει τα μέχρι σήμερα χαρακτηριστικά της αναλογικής παρέχοντας εξελιγμένες δυνατότητες πρόσβασης σε οπτικοακουστική πληροφορία, μεταδιδόμενης μέσω δικτύων ευρείας ζώνης. Κυρίως όμως έρχεται να τις προσδώσει χαρακτηριστικά που μέχρι σήμερα γνωρίσαμε μέσα από την εκρηκτική εξάπλωση του διαδικτύου. Αυτά συνοψίζονται στην εξατομίκευση της πρόσβασης και κατανάλωσης των περιεχομένων από τον χρήστη, την αλληλεπίδραση του χρήστη με το σύστημα και τέλος την ανεξαρτησία από τις χρονικές και χωρικές παραμέτρους της μετάδοσης της πληροφορίας. Η υλοποίηση αυτών των χαρακτηριστικών με χρήση ευρέως αποδεκτών τεχνικών προδιαγραφών ψηφιακής τηλεόρασης και η μελέτη της αποτελεσματικότητας των σχετικών αλγορίθμων και μηχανισμών, αποτέλεσαν τους κύριους στόχους του Ευρωπαϊκού προγράμματος UP-TV(Ubiquitous Personalized TV). Στα πλαίσια αυτά, στην παρούσα εργασία ασχοληθήκαμε με την δημιουργία του προφίλ των χρηστών της ψηφιακής τηλεόρασης και την αυτόματη αφομοίωση από το προφίλ των ιδιαίτερων χαρακτηριστικών που καταγράφονται κατά την αλληλεπίδραση του χρήστη με το σύστημα. Το προφίλ των χρηστών της ψηφιακής τηλεόρασης που κατασκευάσαμε υπακούει στις προδιαγραφές των διεθνών προτύπων για ψηφιακή τηλεόραση(TV Anytime). Το TV Anytime όμως, ενώ καθορίζει κατηγορίες και δομές μεταδεδομένων για το προφίλ των χρηστών, δεν καθορίζει τους τρόπους που μπορούν να χρησιμοποιηθούν τα μεταδεδομένα των χρηστών για να συνδυαστούν με τα μεταδεδομένα των προγραμμάτων της τηλεόρασης. Χρησιμοποιήσαμε δυο διαφορετικούς τρόπους δημιουργίας ερωτήσεων από τις κατηγορίες των μεταδεδομένων των χρηστών και δημιουργήσαμε δομές που τις αναπαριστούν. Κατασκευάσαμε επίσης αλγορίθμους δυναμικής αναπροσαρμογής της πληροφορίας για τις προτιμήσεις των χρηστών. Τέλος διεξήγαμε σειρά πειραμάτων για τον έλεγχο της λειτουργίας του συστήματος και μια πρώτη προσέγγιση για την αξιολόγηση της απόδοσης του συστήματος. Τα πειράματα βασίστηκαν σε δεδομένα προτιμήσεων χρηστών για περίπου 1600 ταινίες με 100000 βαθμολογίες. Κατασκευάσαμε διαφορετικούς αλγορίθμους οι οποίοι από τις βαθμολογίες των χρηστών δημιουργούν αυτόματα προφίλ για τους χρήστες, που εκφράζουν τις προτιμήσεις τους. Χρησιμοποιήσαμε τα προφίλ των χρηστών

και τις εκφράσεις ερωτήσεων που συνεπάγονται για να προβλέψουμε το πόσο θα αρέσουν στο χρήστη κάποια από τα προγράμματα που δεν έχει δει. Δείξαμε ότι τα αποτελέσματα τις πρόβλεψης (με μέσο λάθος ή precision/recall) ήταν ικανοποιητικά σε κάποιους από τους αλγορίθμους και καλά συγκριτικά με την απόδοση των αλγορίθμων συνεργατικής αναζήτησης που μπορεί να εφαρμοστούν μόνο πάνω από αρχιτεκτονική client-server.

Ευχαριστήρια

Θα ήθελα να εκφράσω τις ιδιαίτερες ευχαριστίες μου στον επιβλέποντα καθηγητή μου κ. Σ. Χριστοδουλάκη για την πολύτιμη συμβολή του στην ολοκλήρωση αυτής της εργασίας. Συμβολή που ξεκινά από την καθοδήγησή του και φτάνει μέχρι την εμπειρία και την μοναδική προσωπικότητά του.

Θα ήθελα, επίσης, να ευχαριστήσω τους καθηγητές κκ. Κουμπαράκη Μανώλη και Πετράκη Ευρυπίδη για τον χρόνο που διέθεσαν για την ανάγνωση του κειμένου και τις εποικοδομητικές παρατηρήσεις τους.

Ιδιαίτερα θα ήθελα να ευχαριστήσω τον Μουμουτζή Νεκτάριο για την επίβλεψή του και την πραγματική βοήθεια που μου πρόσφερε όλον αυτόν το καιρό με την ουσιαστική του εμπειρία και την άριστη συνεργασία. Ευχαριστώ επίσης των Παππά Νίκο για την δικιά του συμβολή με την τεχνική του κατάρτιση και την εποικοδομητική συνεργασία του.

Τέλος θα ήθελα να ευχαριστήσω όλη την ομάδα των μελών του εργαστηρίου για την αρμονική μας συνύπαρξη στον ίδιο χώρο καθώς και την συνεργασία τους.

Αφιέρωση

στους γονείς μου

Κεφάλαιο 1: *Εισαγωγή*

Γενικά

Ο σύγχρονος οπτικοακουστικός τομέας χαρακτηρίζεται από μια τεράστια ποικιλία και όγκο προσφερόμενης πληροφορίας με χιλιάδες κανάλια να μεταδίδουν επί 24ώρου βάσεως τηλεοπτικό ή/και ακουστικό περιεχόμενο. Η συνεχής διεύρυνση των προσφερόμενων επιλογών είναι μια διαρκώς αυξανόμενη τάση. Σύντομα ο κάθε τηλεθεατής θα έχει να διαλέξει ανάμεσα σε χιλιάδες κανάλια πράγμα που σημαίνει ότι και μόνο η απλή επισκόπηση των προσφερόμενων προγραμμάτων με τη μορφή τηλεοπτικών οδηγών θα απαιτεί δεκάδες ώρες. Ο σύγχρονος τηλεθεατής επιθυμεί να τιθασεύσει αυτό το χείμαρρο πληροφοριών με τον πιο αποτελεσματικό τρόπο ώστε να μπορεί ανά πάσα στιγμή να επιλέγει την τηλεοπτική εκπομπή που ταιριάζει περισσότερο με τα προσωπικά ή οικογενειακά του ενδιαφέροντα. Αυτή την ανάγκη θα περιγράφουμε με τη φράση «εξατομικευμένη πρόσβαση στην πληροφορία». Μέχρι σήμερα, αυτή η ανάγκη, δεν καλύπτονταν ή καλύτερα καλύπτονταν με τρόπους μη συστηματικούς που βασίζονταν στη δημοσίευση τηλεοπτικών οδηγών τους οποίους ο τηλεθεατής ήταν υποχρεωμένος να μελετά πριν κάνει την επιλογή του σταθμού της αρεσκείας του.

Επιπλέον, ο τηλεθεατής παραμένει παθητικός δέκτης της εκπεμπόμενης πληροφορίας. Αυτό έχει να κάνει με τους τεχνικούς περιορισμούς που το ίδιο το μέσο της τηλεόρασης επιβάλει. Επομένως μια βασική ανάγκη που παραμένει ανικανοποίητη είναι η αμφίδρομη επικοινωνία ώστε ο τηλεθεατής να πάψει να είναι μόνο δέκτης και να μπορεί να λειτουργήσει και ως πομπός πληροφοριών εξασφαλίζοντας, ανάμεσα στα άλλα, την προσαρμογή των τηλεοπτικών προγραμμάτων στα ιδιαίτερα ενδιαφέροντα και προτιμήσεις του.

Συνοψίζοντας, η γνωστή σε όλους αναλογική τηλεόραση (και κατ' επέκταση ο αναλογικός οπτικοακουστικός τομέας) δεν έχει καταφέρει να ικανοποιήσει δύο βασικές ανάγκες των χρηστών του: την εξατομίκευση και την αλληλεπιδραστικότητα. Και ναι μεν παλαιότερα αυτό δεν ήταν τόσο κρίσιμο, στο σημερινό όμως περιβάλλον αποτελεί όρο επιβίωσης της οπτικοακουστικής βιομηχανίας. Θα έλεγε κανείς, ότι δίχως την ικανοποίηση των βασικών αυτών αναγκών, είναι αδύνατον να παρασχεθούν πλέον οπτικοακουστικές υπηρεσίες.

Η λύση θα ήταν εξαιρετικά δυσχερής, αν όχι αδύνατη, χωρίς τη χρήση ενός νέου τεχνολογικού υπόβαθρου που βασίζεται στη σύγκλιση της πληροφορικής, των τηλεπικοινωνιών και του οπτικοακουστικού τομέα Η σύγκλιση αυτή αποτελεί ένα από τα σημαντικότερα επιτεύγματα της σύγχρονης ψηφιακής τεχνολογίας και καρπός της είναι η ψηφιακή τηλεόραση η οποία προσφέρει ήδη ένα νέο μέσο μετάδοσης πληροφοριών το οποίο συνδυάζει τις ιδιότητες της αναλογικής τηλεόρασης με τα κύρια χαρακτηριστικά που περιγράψαμε ως ανάγκες: εξατομικευμένη πρόσβαση ανεξάρτητη από τους χωρικούς και χρονικούς περιορισμούς μετάδοσης και ισχυροί μηχανισμοί αλληλεπίδρασης χρήστη – συστήματος.

Ο καλύτερος ορισμός για την ψηφιακή τηλεόραση είναι μέσα από τα χαρακτηριστικά της, τα σημαντικότερα από τα οποία είναι τα εξής:

- Βελτιωμένη ποιότητα εικόνας και ήχου ακόμα και μέσα από τις υπάρχουσες συσκευές.
- Μεγάλο εύρος ζώνης που συνεπάγεται περισσότερα κανάλια και περισσότερα θεματικά κανάλια.
- Αλληλεπιδραστικές πολυμεσικές υπηρεσίες.

Η ψηφιακή τηλεόραση βασίζεται κατά κύριο λόγο στην τεχνολογία των ηλεκτρονικών υπολογιστών. Η τηλεοπτική συσκευή, ενσωματώνει έναν πλήρη ηλεκτρονικό υπολογιστή καθώς και μεγάλη δευτερεύουσα μνήμη που επιτρέπει την καταγραφή των τηλεοπτικών προγραμμάτων που ο χρήστης επιθυμεί. Όλες οι λειτουργίες της τηλεοπτικής συσκευής, υλοποιούνται με κατάλληλο λογισμικό. Επομένως και οι λειτουργίες που αφορούν την εξατομίκευση, θα πρέπει και αυτές να μπορούν να υλοποιηθούν ως λογισμικό το οποίο χρησιμοποιώντας κατάλληλα δεδομένα θα μπορεί να φιλτράρει τα τηλεοπτικά προγράμματα και να επιλέγει αυτά που ταιριάζουν με τα ενδιαφέροντα και τις προτιμήσεις του τηλεθεατή. Για να καταστεί αυτό δυνατό θα πρέπει να υπάρχουν αυτά τα «κατάλληλα

δεδομένα», δηλ. δεδομένα που θα περιγράφουν τα χαρακτηριστικά γνωρίσματα των τηλεοπτικών προγραμμάτων και τα ενδιαφέροντα των τηλεθεατών. Αυτά τα ειδικού τύπου δεδομένα, ονομάζονται «μεταδεδομένα». Με τον όρο αυτό σήμερα εννοούμε όλη την πληροφορία που στόχος της είναι να περιγράψει τα ίδια τα δεδομένα Κύριος στόχος των μεταδεδομένων είναι η σημασιολογική περιγραφή των δεδομένων και σε ανώτερο επίπεδο της ίδιας της πληροφορίας. Μέσω των μεταδεδομένων η πληροφορία αποκτά την δυνατότητα να φέρει μαζί της και όλη την σημασιολογική πληροφορία που την χαρακτηρίζει. Για να είναι αυτό εφικτό πρέπει όλες οι οντότητες που αποτελούν στοιχεία ενός συστήματος να μπορούν να περιγραφούν τόσο από την πλευρά του ορισμού τους σαν αντικείμενα όσο και από τη πλευρά της περιγραφής τους σαν σημασιολογικές οντότητες.

Στην περίπτωση της ψηφιακής τηλεόρασης οι βασικές οντότητες που μας ενδιαφέρουν είναι από τη μια πλευρά ο τελικός χρήστης και από την άλλη πλευρά το διαθέσιμο οπτικοακουστικό περιεχόμενο. Το οπτικοακουστικό περιεχόμενο περιγράφεται σε επίπεδο τηλεοπτικών προγραμμάτων και χαρακτηρίζεται από την πληροφορία που αναφέρεται στην προέλευσή, την δημιουργία και την κατηγοριοποίηση του. Από την άλλη πλευρά ο χρήστης αλλά και οι ενέργειές του περιγράφονται σε επίπεδο ενδιαφερόντων και προτιμήσεών του. Αυτά είναι και τα απαραίτητα στοιχεία για την εξατομίκευση των παρεχόμενων υπηρεσιών δίνοντας τη δυνατότητα για πρόσβαση στο περιεχόμενο που ενδιαφέρει περισσότερο το χρήστη, την στιγμή και στο σημείο που προτιμά.

Μοντέλα Μεταδεδομένων

Από τη στιγμή που αναγνωρίστηκαν οι παραπάνω ανάγκες και με τη ραγδαία εξάπλωση τέτοιων συστημάτων ήταν απαραίτητη και η ανάπτυξη των αντίστοιχων προτύπων μοντέλων μεταδεδομένων. Η χρήση τέτοιων προτύπων κάνει εφικτή την διαφανή πρόσβαση στα περιεχόμενα διαφορετικών συστημάτων, ανεξάρτητα από την εσωτερική οργάνωση των δεδομένων τους, επιτρέποντας έτσι την διαλειτουργικότητα των συστημάτων και των εφαρμογών τους. Τα μοντέλα μεταδεδομένων ορίζουν σύνολα από χαρακτηριστικά που περιγράφουν την οπτικοακουστική πληροφορία με τρόπο που να καλύπτουν συγκεκριμένες ανάγκες διαχείρισης και ανάκτησής της.

Η πιο έγκυρη και πλήρης προσέγγιση στην μοντελοποίηση μεταδεδομένων πολυμέσων σήμερα, είναι το διεθνές πρότυπο MPEG-7. Παράλληλα όμως γίνονται και προσπάθειες τυποποίησης σχημάτων μεταδεδομένων που αφορούν πιο εξειδικευμένες χρήσεις του οπτικοακουστικού υλικού. Στην περίπτωση της ψηφιακής τηλεόρασης το πιο σημαντικό είναι το σχήμα μεταδεδομένων που προέρχεται από το TVAnytime Forum(TVA). Το TVAF έχει σαν στόχο όχι μόνο την τυποποίηση των μεταδεδομένων αλλά και την σχεδίαση και υλοποίηση ολοκληρωμένων συστημάτων τα οποία να παρέχουν υπηρεσίες σχετικές με οπτικοακουστικό υλικό ευρείας ζώνης εκμεταλλευόμενα την ύπαρξη αποθηκευτικών μέσων μεγάλης χωρητικότητας που βρίσκονται εγκατεστημένα σε συσκευές ψηφιακής τηλεόρασης. Το μοντέλο του TVA σε πολλά σημεία ενσωματώνει περιγραφές του MPEG-7 αλλά οι βασικές του οντότητες είναι διαφορετικές προκειμένου να καλύπτουν τις ανάγκες που περιγράψαμε. Οι απαιτήσεις που θέτει και οι ιδέες πίσω από την ανάπτυξή του αποτελούν βάση για τις τεχνολογίες που αναπτύσσονται στο πρόγραμμα UP-TV στα πλαίσια του οποίου εκπονήθηκε η παρούσα διπλωματική εργασία.

Το πρόγραμμα UP-TV

Το πρόγραμμα UP-TV είναι ένα ευρωπαϊκό ερευνητικό πρόγραμμα που αποσκοπεί στην ανάπτυξη τεχνολογίας που θα παρέχει εξελιγμένες δυνατότητες πρόσβασης σε οπτικοακουστική πληροφορία, μεταδιδόμενης μέσω δικτύων ευρείας ζώνης. Η πρόσβαση στην πληροφορία είναι επιθυμητό να έχει τα ακόλουθα χαρακτηριστικά:

- Εξατομίκευση (personalization) : Ο χρήστης του συστήματος θα έχει την δυνατότητα προσαρμογής της πρόσβασης και κατανάλωσης των περιεχομένων βάσει των προτιμήσεών του.
- Αλληλεπιδραστικότητα (interactivity) : Ο χρήστης θα έχει την δυνατότητα αλληλεπίδρασης με το σύστημα για την εκτέλεση ερωτήσεων αναζήτησης καθώς και αυτόματης οργάνωσης και δεικτοδότησης του περιεχομένου του συστήματος.
- Ανεξαρτησία από τις χρονικές και χωρικές παραμέτρους μετάδοσης της πληροφορίας(ubiquity).

Βάση των συστημάτων που αναπτύσσονται στα πλαίσια του προγράμματος είναι η αρχιτεκτονική που προτείνεται από το TVA. Βασικό δομικό στοιχείο της αρχιτεκτονικής είναι τα Προσωπικά Συστήματα Καταγραφής Βίντεο (PVRs) τα οποία εκμεταλλεύονται την διαρκώς αυξανόμενη χωρητικότητα των αποθηκευτικών μέσων, καθώς και τις προηγμένες τεχνολογίες κωδικοποίησης, για να παρέχουν βασικές δυνατότητες εξατομικευμένης πρόσβασης σε τηλεοπτικές μεταδόσεις ανεξάρτητα από τον χρόνο, αποθηκεύοντας για μετέπειτα χρήση ολόκληρα τηλεοπτικά προγράμματα ή τμήματα αυτών, που ενδιαφέρουν τον χρήστη. Ο χρήστης μπορεί μέσω της συσκευής του να πλοηγηθεί στο περιεχόμενο Ηλεκτρονικών Οδηγών Προγραμμάτων (Electronic Program Guides - EPGs) και να εντοπίσει ενδιαφέροντα αντικείμενα, τα οποία η συσκευή αναλαμβάνει στη συνέχεια να αποθηκεύσει.

Το πρόγραμμα UP-TV επιδιώκει να επεκτείνει την λειτουργικότητά του PVR, εισάγοντας δίκτυα κατανεμημένων εξυπηρετητών, οι οποίοι θα παρέχουν υψηλότερου επιπέδου υπηρεσίες. Το σύστημα θα είναι σε θέση να εντοπίζει τις μεταδόσεις που ενδιαφέρουν το χρήστη και να κάνει προτάσεις πάνω σε αυτές, μέσω σύνθετων προφίλ προτιμήσεων και αντίστοιχων, πλούσιων σε πληροφορία μεταδεδομένων. Τα επιλεγμένα αντικείμενα θα αποθηκεύονται στο δίσκο με τρόπο ώστε να είναι «κοντά» στον τελικό χρήστη, ενώ το

σύστημα σχεδιάστηκε για να υποστηρίζει χρήστες που αλλάζουν γεωγραφική θέση και που επιθυμούν να έχουν γρήγορη πρόσβαση στο υλικό που τους ενδιαφέρει από τον εκάστοτε τόπο διαμονής τους.

Στόχος Εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι **η μελέτη του μοντέλου προφίλ του MPEG7/TVA και η υλοποίηση μηχανισμών αυτόματης κατασκευής και αναπροσαρμογής του**. Λέγοντας προφίλ αναφερόμαστε στο σύνολο των χαρακτηριστικών που αναφέρονται στα ενδιαφέροντα και τις προτιμήσεις των χρηστών. Σε ένα περιβάλλον ψηφιακής τηλεόρασης, ο χρήστης μπορεί να μην είναι σε θέση να προσδιορίσει ο ίδιος άμεσα το προφίλ του και ιδιαίτερα όταν το μοντέλο περιγραφής του προφίλ είναι πολύπλοκο όπως συμβαίνει στην περίπτωση του προτύπου MPEG-7. Για τις περιπτώσεις αυτές το σύστημα θα πρέπει να είναι σε θέση, χρησιμοποιώντας την ιστορία χρήσης, να εκτιμά την προτίμηση του χρήστη για συγκεκριμένα προγράμματα και στη συνέχεια να κατασκευάζει αυτόματα(ή να αναπροσαρμόζει) το προφίλ του χρησιμοποιώντας τα μεταδεδομένα των προγραμμάτων που αξιολογήθηκαν για τον συγκεκριμένο χρήστη.

Page: 7

Ειδικότεροι στόχοι της διπλωματικής εργασίας είναι:

1. Η κατασκευή αυτόματων μηχανισμών εκτίμησης του ενδιαφέροντος του χρήστη για τα προγράμματα που έχει παρακολουθήσει. Η εκτίμηση θα γίνεται με διάφορα κριτήρια: είδος ενεργειών που εκτέλεσε στα προγράμματα, ποσοστό χρόνου παρακολούθησης του προγράμματος σε σχέση με τη συνολική του διάρκεια κ.λ.π.
2. Κατασκευή μηχανισμών συναγωγής και αναπροσαρμογής του προφίλ με βάση την εκτίμηση που έγινε για τα προγράμματα που παρακολούθησε ο χρήστης. Το προφίλ θα περιγράφεται με τους δεδομένους μηχανισμούς (Filtering And Search Preferences) του TVA / MPEG7.
3. Συγκριτική αξιολόγηση των εναλλακτικών μηχανισμών κατασκευής του προφίλ (πειράματα).

Χρησιμοποιούμενες Τεχνολογίες

Το σύστημα που αναπτύχθηκε είναι βασισμένο σε TVA συμβατή βάση δεδομένων η οποία είναι υλοποιημένη σε MySQL 4.0. Το τμήμα της κατασκευής του προφίλ καθώς και των πειραμάτων έχει υλοποιηθεί σε MySQL ενώ η αναπροσαρμογή των προφίλ είναι υλοποιημένη σε java (jdk 1.4). Επίσης για τη διαχείριση XML εγγράφων και δεδομένων από τη βάση σύμφωνα με το σχήμα του TVA έγινε χρήση των δομών του Breeze XML Binder.

Σχετιζόμενη Εργασία

Η παρούσα εργασία κινείται στην περιοχή των τεχνολογιών που αναπτύσσονται για να υποστηρίξουν την ψηφιακή τηλεόραση. Το κύριο πρόβλημα που πρέπει να αντιμετωπιστεί είναι αυτό του πολύ μεγάλου αριθμού προγραμμάτων που είναι διαθέσιμα ανα πάσα στιγμή στο χρήστη. Το πρόβλημα είναι ανάλογο του τεράστιου όγκου πληροφοριών που είναι διαθέσιμες στο διαδίκτυο, για αυτό και ήδη ανεπτυγμένες τεχνολογίες βρίσκουν εκ νέου εφαρμογή στην περίπτωση της ψηφιακής τηλεόρασης. Κύριος στόχος είναι η εξατομίκευση της πρόσβασης στα δεδομένα της ψηφιακής τηλεόραση από την μια πλευρά και η ανάπτυξη συστημάτων που θα επιτρέπουν εξατομικευμένες προτάσεις προγραμμάτων για την διευκόλυνση του κάθε χρήστη.

Για τους σκοπούς αυτούς πολύ σημαντικό ρόλο παίζουν τα ενδιαφέροντα κάθε χρήστη τα οποία στοιχειοθετούν το προφίλ του. Στη σχετική δουλειά που έχει γίνει μέχρι σήμερα γίνονται προσπάθειες για αυτόματη εξαγωγή των ενδιαφερόντων κάθε χρήστη χωρίς να απαιτείται η παρέμβασή του. Η δουλειά σε αυτόν τον τομέα προέρχεται κυρίως από την περιοχή των εφαρμογών του διαδικτύου και βασικό εργαλείο για αυτόν το σκοπό είναι η καταγραφή των ενεργειών του χρήστη κατά την αλληλεπίδρασή του με το σύστημα. Σε αυτήν την κατεύθυνση της μελέτης των ενεργειών του χρήστη και τις ένδειξης που μπορούν να δώσουν για τα ενδιαφέροντα του κινείται το [1]. Η διάρκεια της αλληλεπίδρασης του χρήστη, η κινητικότητά του σε κάθε σελίδα είναι βασικά στοιχεία για την εξαγωγή συμπερασμάτων που αφορούν στα ενδιαφέροντα του χρήστη. Σε μια πιο συστηματική ομαδοποίηση και κατηγοριοποίηση των ενεργειών του χρήστη οδηγούνται τα [2], [3] μέσα από την παρουσίαση προηγούμενων εργασιών τόσο στις περιοχές του φιλτραρίσματος και της αναζήτησης όσο και ειδικά στην μελέτη συστημάτων συλλογής ρητών και μη ρητών προτιμήσεων των χρηστών. Στην εκτίμηση του ενδιαφέροντος του χρήστη για μια ιστοσελίδα βάση των ενεργειών του, κινείται και το [4] ενώ τέλος στο [5] η καταγραφή της αλληλεπίδρασης του χρήστη με το σύστημα εξετάζεται από την πλευρά της εξόρυξης δεδομένων(Data Mining).

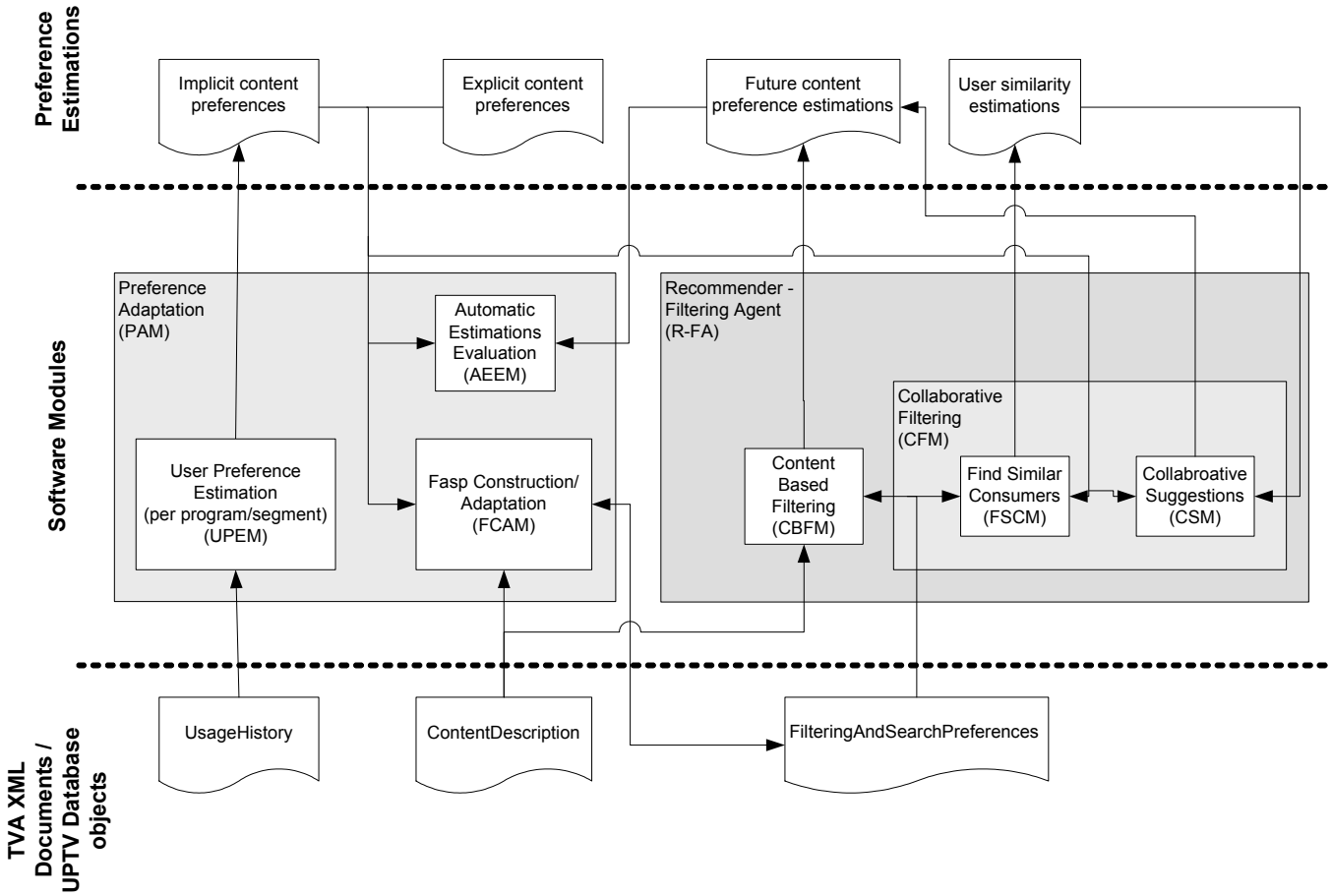
Για την ολοκλήρωση ενός συστήματος και την εκμετάλλευση των ενδιαφερόντων του χρήστη χρειάζεται η συστηματική καταγραφή τους σε αυτό που ονομάζουμε προφίλ του χρήστη και περιέχει όλη την πληροφορία που τον χαρακτηρίζει. Με την μοντελοποίηση του προφίλ του χρήστη ασχολείται το [6] ενώ στο [7] ρητές και μη ρητές προτιμήσεις του

χρήστη δημιουργούν το προφίλ του χρήστη που με τη βοήθεια συνδυαστικού φιλτραρίσματος μπορούν να χρησιμοποιηθούν για ένα ολοκληρωμένο σύστημα προτάσεων.

Τα συστήματα προτάσεων είναι μια κατεύθυνση στην οποία κινούνται αρκετά συστήματα στην περιοχή της ψηφιακής τηλεόρασης. Κύριος στόχος αυτών των συστημάτων είναι η δημιουργία εξατομικευμένων ηλεκτρονικών τηλεοπτικών οδηγών(EPGs). Τέτοια συστήματα είναι τα [8], [9]. Στη δημιουργία συστήματος προτάσεων στοχεύει και το [10] με τη δημιουργία προφίλ για του χρήστες και τη χρήση συνδυαστικού φιλτραρίσματος ενώ το [11] εξυπηρετεί τους ίδιος σκοπούς βασιζόμενο σε εξειδικευμένους πράκτορες και εφαρμογή με χρήση του TV-Anytime. Τέλος ένα ολοκληρωμένο σύστημα προτάσεων είναι το TV Scout [12], με κύρια μέριμνα την επικοινωνία μέσω κάποιου γραφικού περιβάλλοντος με το χρήστη.

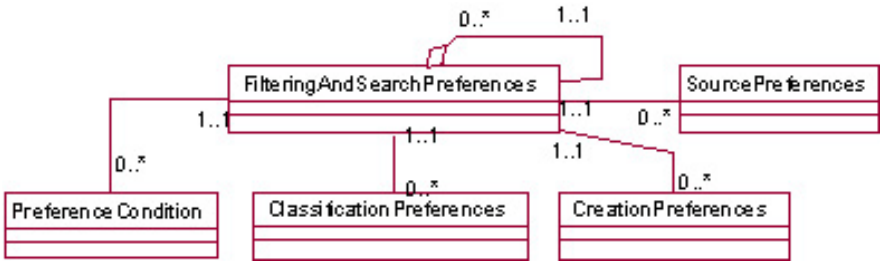
Κεφάλαιο 2: *Αρχιτεκτονική του Συστήματος*

Το τμήμα του συνολικού συστήματος με το οποίο θα ασχοληθούμε είναι αυτό που περιλαμβάνει την διαχείριση των μεταδεδομένων περιγραφής των προτιμήσεων και των ενεργειών των χρηστών από την μια πλευρά και των μεταδεδομένων περιγραφής του οπτικοακουστικού περιεχομένου από την άλλη. Τελικός στόχος είναι η συσχέτιση των προτιμήσεων του χρήστη με τις περιγραφές του περιεχομένου προκειμένου να εντοπίζεται το περιεχόμενο ενδιαφέροντος του χρήστη. Η αρχιτεκτονική του συστήματος φαίνεται στο επόμενο σχήμα:



Σχήμα 1: Συνολική αρχιτεκτονική του συστήματος

Όπως διακρίνουμε το σύστημα αποτελείται από δύο βασικά υποσυστήματα: το PAM και το R-FA. Το πρώτο δέχεται σαν είσοδο την περιγραφή των διαθέσιμων στο σύστημα προγραμμάτων, το περιεχόμενο των οποίων ακολουθεί την περιγραφή μεταδεδομένων του TVA. Στη συνέχεια χρησιμοποιεί τις προτιμήσεις των χρηστών για όσα πρόγραμμα είναι διαθέσιμες προκειμένου να παράγει το προφίλ κάθε χρήστη σε μορφή TVA.. Η μορφή αυτή μπορεί να είναι είτε απλή είτε ιεραρχική όπως περιγράφεται από το παρακάτω σχήμα:



Σχήμα 2: Περιγραφή προτιμήσεων χρηστών σύμφωνα με το TVA

Εδώ θα πρέπει να σημειώσουμε πως οι προτιμήσεις των χρηστών μπορεί να είναι είτε ρητές είτε να προκύπτουν αυτόματα με βάση την το ιστορικό των ενεργειών που καταγράφονται κατά την αλληλεπίδραση του χρήστη με το σύστημα. Επίσης είναι δυνατό ένα νέο προφίλ για κάποιον χρήστη να χρησιμοποιηθεί προκειμένου να γίνει αναπροσαρμογή κάποιου παλιού. Από την άλλη πλευρά το R-FA δέχεται σαν είσοδο τις περιγραφές των προγραμμάτων και τα προφίλ των χρηστών και με «έξυπνους» μηχανισμούς φιλτραρίσματος τα συσχετίζει προκειμένου να κάνει εκτιμήσεις για τις προτιμήσεις των χρηστών πάνω σε προγράμματα τα οποία δεν έχει δει. Ο συσχετισμός αυτός γίνεται με χρήση του p-norm extended Boolean Μοντέλου το οποίο υλοποιείται στο R-FA. Τέλος τα δύο υποσυστήματα συνεργάζονται προκειμένου να γίνεται αξιολόγηση της απόδοσης των διάφορων μηχανισμών καθώς και των εκτιμήσεων του συστήματος.

Όπως ήδη έχουμε αναφέρει διακρίνουμε δυο υποσυστήματα: το υποσύστημα Προσαρμογής Προτιμήσεων (PAM) και το υποσύστημα που είναι υπεύθυνο για της Προτάσεις και το Φιλτράρισμα των προγραμμάτων(R-FA). Στην παρούσα εργασία υλοποιήθηκε το πρώτο υποσύστημα στην αναλυτικότερη περιγραφή του οποίου θα προχωρήσουμε αμέσως τώρα. Τα τμήματα που το αποτελούν είναι τα εξής:

Υποσύστημα Εκτίμησης Προτιμήσεων(UPEM)

Το UPEM περιλαμβάνει μηχανισμούς που δέχονται σαν είσοδο μια ιστορία χρήσης και παράγουν βάσει διαφόρων κριτηρίων, μια εκτίμηση της προτίμησης του χρήστη για κάθε πρόγραμμα που παρακολούθησε (ή απλώς διαχειρίστηκε). Τα κριτήρια που χρησιμοποιούνται για την εκτίμηση αυτή είναι μεταξύ άλλων:

1. Το είδος των ενεργειών που εκτελέστηκαν για το συγκεκριμένο πρόγραμμα.
2. Το ποσοστό του προγράμματος που παρακολούθησε ο χρήστης.

Στην περίπτωση που υπάρχει πληροφορία τμηματοποίησης (segmentation) τα κριτήρια που μπορούν να χρησιμοποιηθούν για την εκτίμηση της προτίμησης του χρήστη για καθένα από αυτά είναι τα εξής:

1. Το είδος των ενεργειών που εκτελέστηκαν για το κάθε τμήμα
2. Το ποσοστό του τμήματος που παρακολούθησε ο χρήστης.
3. Ο εντοπισμός της προτίμησης για ανεξάρτητα διαστήματα του προγράμματος που παρακολούθησε ο χρήστης και ο συσχετισμός τους με την υπάρχουσα πληροφορία τμηματοποίησης.

Υποσύστημα Κατασκευής και Προσαρμογής Προφίλ(FCAM)

Το FCAM υλοποιεί μηχανισμούς για την αυτόματη κατασκευή και αναπροσαρμογή των προτιμήσεων φιλτραρίσματος και αναζήτησης (Filtering And Search Preferences). Οι μηχανισμοί αυτοί, παίρνουν σαν είσοδο μια σειρά προτιμήσεων (ή εκτιμήσεων προτίμησης) του χρήστη για συγκεκριμένα προγράμματα ή τμήματα προγραμμάτων, και χρησιμοποιούν την πληροφορία περιεχομένου (content description) για να κατασκευάσουν μια δομή προτιμήσεων φιλτραρίσματος και αναζήτησης για το χρήστη αποδίδοντας στις παραμέτρους αυτής της δομής κατάλληλες τιμές. Εδώ μπορούμε να διακρίνουμε δύο καίρια ζητήματα:

1. Η δομή των προτιμήσεων φιλτραρίσματος μπορεί να είναι απλή ή ιεραρχική. Αν είναι ιεραρχική απαιτείται ο καθορισμός του χαρακτηριστικού που προσδιορίζει την ιεραρχία (κατά κανόνα το genre). Εναλλακτικά η ιεραρχική δομή μπορεί να βασιστεί στην σχέση εξειδίκευσης μεταξύ απλών δομών προτιμήσεων.
2. Οι παράμετροι μιας δεδομένης δομής που πρέπει να αρχικοποιηθούν για να είναι δυνατόν αυτή η δομή να χρησιμοποιηθεί στην ανάκτηση περιεχομένου. Οι παράμετροι αυτοί αναφέρονται στο σχετικό βάρος με το οποίο κάθε συνιστώσα της

δομής συμμετέχει στην αποτίμηση κάθε προγράμματος ή τμήματος προγράμματος. Αλλάζοντας τις τιμές αυτών των παραμέτρων, διαφοροποιείται η σχετική βαρύτητα των συνιστωσών και προκύπτουν διαφορετικά αποτελέσματα.

Υποσύστημα Αξιολόγησης Εκτιμήσεων(AEEM)

Το AEEM είναι ένα υποσύστημα που αποσκοπεί στην αξιολόγηση των μηχανισμών του UPEM και FCAM, καθώς στην ρύθμιση των παραμέτρων που χρησιμοποιούν οι μηχανισμοί αυτοί. Ειδικότερα:

1. Για την αξιολόγηση των μηχανισμών του UPEM, το AEEM δέχεται σαν είσοδο τις εκτιμήσεις προτίμησης ανά πρόγραμμα που υπολογίζει ο κάθε μηχανισμός του UPEM και συγκρίνει τα αποτελέσματα με ρητές προτιμήσεις που έχουν δώσει κάποιοι χρήστες.
2. Για την αξιολόγηση των μηχανισμών του FCAM, το AEEM δέχεται σαν είσοδο τις εκτιμήσεις προτίμησης για μελλοντικά προγράμματα ή τμήματα προγραμμάτων που προκύπτουν από τον R-FA. Στη συνέχεια, και αφού κάποιοι επιλεγμένοι χρήστες παρακολουθήσουν τα προτεινόμενα προγράμματα και τα αξιολογήσουν ρητά, το AEEM συγκρίνει αυτές τις ρητές προτιμήσεις με τις εκτιμήσεις του R-FA για να αξιολογήσει την ακρίβεια των προτιμήσεων φιλτραρίσματος και αναζήτησης που κατασκεύασε το FCAM τόσο ως προς εναλλακτικές δομές αυτών όσο και ως προς διαφορετικούς τρόπους υπολογισμού των παραμέτρων κάθε δομής.

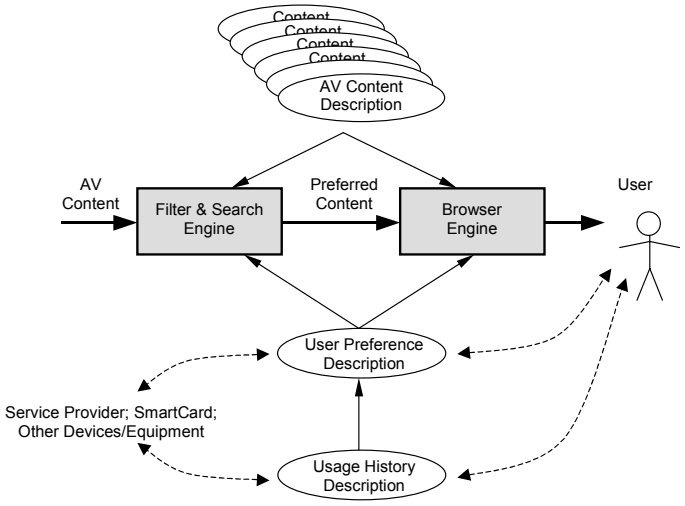
Στα πλαίσια της διπλωματικής εκτός από την ανάπτυξη των παραπάνω μονάδων του συστήματος **σχεδιάστηκαν και πραγματοποιήθηκαν** πειράματα για την αξιολόγηση των μηχανισμών των μονάδων UPEM και FCAM. Ειδικότερα, στα πειράματα αυτά θα χρησιμοποιούν ρητές προτιμήσεις χρηστών για προγράμματα. Με βάση αυτές τις ρητές εκτιμήσεις, το FCAM για κάθε χρήστη θα κατασκευάσει (ή θα αναθεωρήσει) το προφίλ του με όλους τους δυνατούς τρόπους που προκύπτουν από τους μηχανισμούς που υλοποιεί. Ένα νέο σύνολο προγραμμάτων θα ανακτηθεί με βάση τα νέα εναλλακτικά προφίλ και θα

επιστραφεί στο χρήστη μαζί με μια εκτίμησης της προτίμησής του για αυτά. Στη συνέχεια οι ρητές προτιμήσεις του χρήστη για τα προγράμματα θα συγκριθούν με τις εκτιμήσεις που έδωσε ο μηχανισμός του R-FA για καθένα από τα εναλλακτικά προφίλ από το AEEM ώστε να εντοπιστούν τα προφίλ που έδωσαν τα πλησιέστερα προς τις ρητές προτιμήσεις του χρήστη αποτελέσματα. Με τρόπο αυτό θα αναδειχθούν οι μηχανισμοί που έχουν την καλύτερη απόδοση.

Κεφάλαιο 3: *Περιγραφή Μεταδεδομένων*

Ο όρος «μεταδεδομένα» χρησιμοποιείται, όπως ήδη έχουμε αναφέρει, για να περιγράψει την πληροφορία για τα ίδια τα δεδομένα. Στην περίπτωσή μας εκφράζει την οποιαδήποτε διαθέσιμη πληροφορία για το περιεχόμενο της ψηφιακής τηλεόρασης δηλαδή τους τίτλους των προγραμμάτων, τους ηθοποιούς, τις περιλήψεις κ.τ.λ. Τέτοιου είδους μεταδεδομένα αποτελούν την αναλυτική περιγραφή των χαρακτηριστικών κάθε αντικείμενου και άρα το μέσο διάκρισής των αντικειμένων μεταξύ τους. Έτσι παρέχουν στον χρήστη την δυνατότητα για αναζήτηση, περιήγηση και διαχείριση του διαθέσιμου περιεχομένου, λειτουργίες που θα τον οδηγήσουν στην τελική επιλογή του περιεχομένου προς κατανάλωση. Επιπρόσθετα και δεδομένου ότι το σύστημα είναι τόσο αλληλεπιδραστικό όσο και αμφίδρομο υπάρχουν ανάλογα μεταδομένα και από την πλευρά του χρήστη για την περιγραφή του ίδιου του χρήστη και συγκεκριμένα των ενδιαφερόντων και των προτιμήσεών του. Τέτοιου είδους πληροφορία είναι αξιοποιήσιμη από την πλευρά του συστήματος το οποίο αποκτά με αυτόν τον τρόπο την δυνατότητα να εντοπίζει αυτόματα, πιθανώς ενδιαφέρον για τον χρήστη περιεχόμενο και να του κάνει προτάσεις για τα προγράμματα που το αντιπροσωπεύουν.

Αναλυτικότερα οι βασικές έννοιες που χρησιμοποιούνται φαίνονται στο παρακάτω σχήμα.



Σχήμα 3: Υποστηριζόμενη από το σύστημα λειτουργικότητα

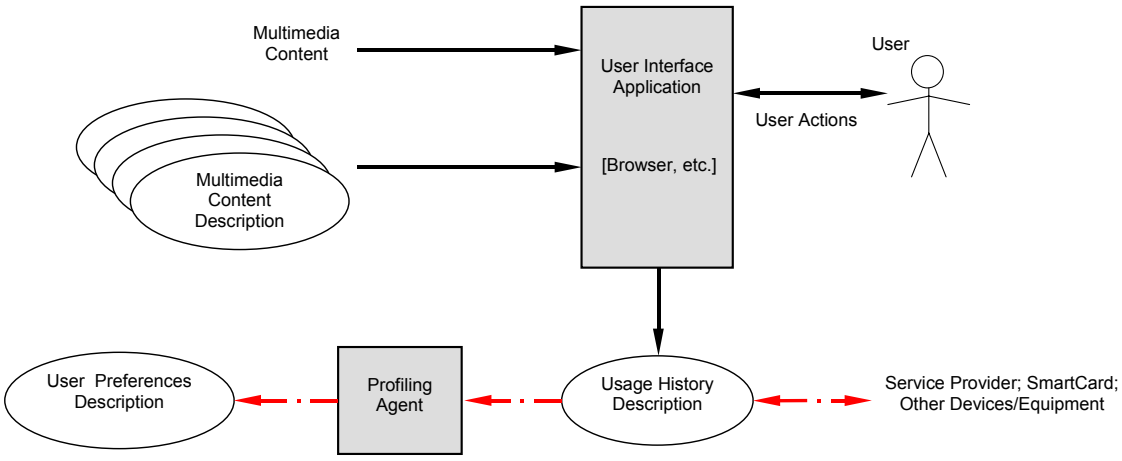
Ο χρήστης αλληλεπιδρά με το περιεχόμενο χρησιμοποιώντας ένα σύστημα πολυμέσων. Το σύστημα χρησιμοποιείται για τον εντοπισμό του περιεχομένου και για την κατανάλωσή του. Η περιγραφή του περιεχομένου είναι διαθέσιμη στο σύστημα προκειμένου να παρέχεται η δυνατότητα για αποδοτικό φιλτράρισμα, αναζήτηση και περιήγηση στο περιεχόμενο. Η περιγραφή των προτιμήσεων του χρήστη είναι επίσης διαθέσιμη ώστε να είναι εφικτό το εξατομικευμένο φιλτράρισμα, αναζήτηση και περιήγηση στο περιεχόμενο. Οι προτιμήσεις του χρήστη χρησιμοποιούνται ώστε να εντοπίζεται το ενδιαφέρον, για αυτόν, περιεχόμενο και να του παρέχεται με τον προτιμητέο τρόπο παρουσίασης. Το σύστημα δύναται επίσης να παράγει μια περιγραφή της ιστορίας χρήσης του χρήστη βασισμένη στο ιστορικό των αλληλεπιδράσεων του χρήστη με το περιεχόμενο. Αυτή η ιστορία χρήσης μπορεί να χρησιμοποιηθεί για την απευθείας αντιστοίχησή της σε προτιμήσεις του χρήστη.

Μεταδεδομένα Καταναλωτών

Σχήμα Περιγραφής Ιστορίας Χρήσης

Το πρώτο σχήμα που θα μας απασχολήσει εδώ είναι το σχήμα περιγραφής της πληροφορίας για την ιστορία χρήσης του χρήστη που συγκεντρώνεται για εκτεταμένες περιόδους αλληλεπίδρασης του χρήστη με το σύστημα. Τα δεδομένα που συλλέγονται αποτελούνται από λίστες ενεργειών του χρήστη οι οποίες στη συνέχεια μπορούν με αυτόματες μεθόδους ανάλυσης να οδηγήσουν στις προτιμήσεις του χρήστη.

Ένα γενικό διάγραμμα μιας εφαρμογής που λαμβάνει υπόψη τις ενέργειες του χρήστη και τις περιγραφές του περιεχομένου και παράγει σαν έξοδο οργανωμένες περιγραφές του ιστορικού κατανάλωσης του περιεχομένου φαίνεται στο παρακάτω σχήμα.

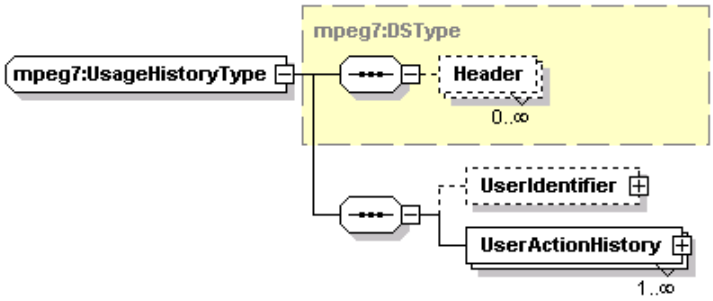


Σχήμα 4: Χρήση ιστορικού ενεργειών χρήστη για εξαγωγή προτιμήσεων και κατασκευή προφίλ

Κάποια τυποποιημένη δομή είναι απαραίτητη όταν απαιτείται η ανταλλαγή τέτοιας πληροφορίας προκειμένου να εξασφαλίζεται η διαλειτουργικότητα μεταξύ διαφορετικών συσκευών και πλατφόρμων. Το σχήμα που έχει οριστεί από το TVA-Forum για αυτόν τον σκοπό βρίσκεται στο ISO/IEC 15938-5. Ακολουθεί η αναλυτική περιγραφή του σχήματος.

Ιστορία Χρήσης

Η ιστορία χρήσης είναι το αντικείμενο του υψηλότερου επιπέδου και περιγράφει την κατανάλωση του οπτικοακουστικού περιεχομένου για έναν χρήστη, σαν λίστα από ενέργειες που διεξήχθησαν σε μια συγκεκριμένη χρονική περίοδο.

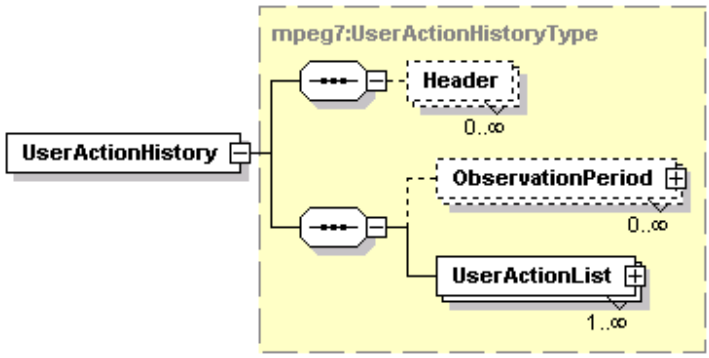


Σχήμα 5: Σχήμα Ιστορίας Χρήσης

Ένα στιγμιότυπο ιστορικού χρήσης περιέχει ένα στοιχείο αναγνωριστικό του χρήστη(UserIdentifier), το οποίο προσδιορίζει τον χρήστη ή την ομάδα χρηστών στους οποίους αναφέρεται η συγκεκριμένη πληροφορία κατανάλωσης. Το ιστορικό χρήσης προσδιορίζεται από το σχήμα της *Ιστορίας Ενεργειών του Χρήστη*(UserActionHistory). Στην αμέσως επόμενη ενότητα περιγράφονται όλα τα αντικείμενα του σχήματος.

Ιστορία Ενεργειών Χρήστη

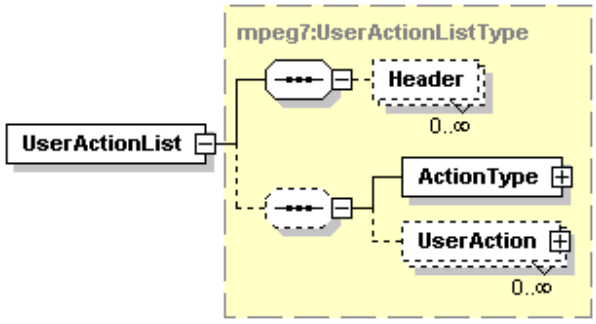
Η ιστορία ενεργειών του χρήστη περιγράφει πολλαπλές λίστες ενεργειών κάθε μια από τις οποίες αναφέρεται σε καταγραφές ενεργειών ενός μόνο τύπου όπως «αναπαραγωγή», «εγγραφή», «επισκόπηση» κ.τ.λ. Όπως παρατηρούμε υπάρχουν μία οι περισσότερες λίστες ενεργειών ενώ υπάρχει και ένα αντικείμενο(ObservationPeriod) για τον προσδιορισμό της περιόδου καταγραφής των ενεργειών.



Σχήμα 6: Σχήμα Ιστορικού Ενεργειών Χρήστη

Λίστα Ενεργειών Χρήστη

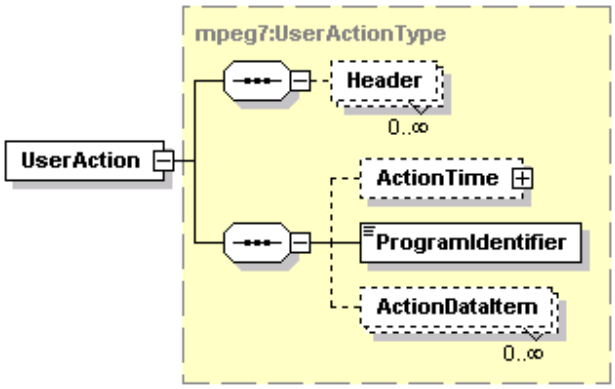
Η λίστα ενεργειών του χρήστη ορίζει μια δομημένη λίστα από αντικείμενα ενεργειών και είναι οργανωμένη σύμφωνα με τους τύπους των ενεργειών. Κάθε ενέργεια είναι συσχετισμένη με ένα μόνο πρόγραμμα η αντικείμενο περιεχομένου.



Σχήμα 7: Σχήμα Λίστας Ενεργειών

Ενέργεια Χρήστη

Η Ενέργεια Χρήστη παρέχει αναλυτική πληροφορία για κάθε μεμονωμένη ενέργεια του χρήστη όπως ο χρόνος που συνέβη, η διάρκειά της, το σχετιζόμενο πρόγραμμα με αυτήν, η τοποθεσία του προγράμματος και αναφορές στη σχετιζόμενη περιγραφή του περιεχομένου αλλά και το αντίστοιχο υλικό.

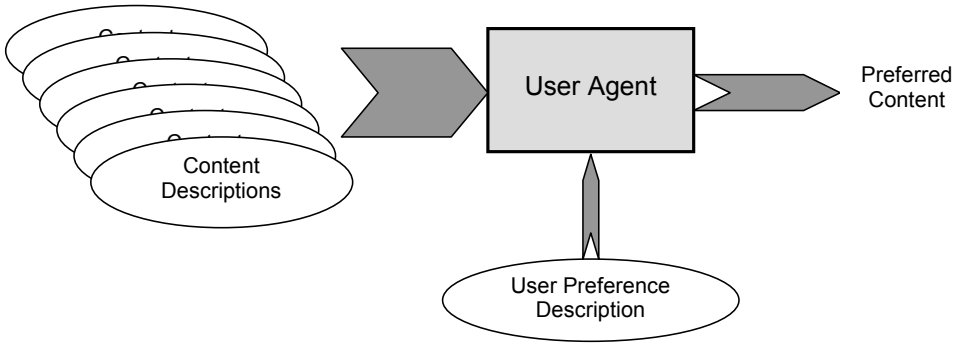


Σχήμα 8: Σχήμα Ενεργειών Χρήστη

Σχήμα Περιγραφής Προτιμήσεων Χρηστών

Η ενότητα αυτή περιέχει τα σχήματα περιγραφής που διευκολύνουν την περιγραφή των προτιμήσεων του χρήστη σχετικά με το υπό κατανάλωση οπτικοακουστικό υλικό. Οι προτιμήσεις των χρηστών μπορούν να συσχετιστούν με τις περιγραφές του περιεχομένου για να εξυπηρετηθεί η αναζήτηση, το φιλτράρισμα και τελικά η επιλογή του επιθυμητού περιεχομένου. Η αντιστοιχία μεταξύ των προτιμήσεων των χρηστών και της περιγραφή του περιεχομένου διευκολύνει την αποδοτική εξατομίκευση της πρόσβασης και κατανάλωσης του περιεχομένου.

Ένα γενικό μοντέλο χρήσης παρουσιάζεται στο παρακάτω σχήμα:



Σχήμα 9: Λειτουργία συσχέτισης περιγραφών δεδομένων με προτιμήσεις χρηστών για εντοπισμό ενδιαφέροντος περιεχομένου.

Το σύστημα παίρνει σαν είσοδο περιγραφές του περιεχομένου και προτιμήσεις χρηστών και παράγει φιλτραρισμένη έξοδο καθορίζοντας τα αντικείμενα εκείνα που ταιριάζουν στις προτιμήσεις του χρήστη. Σε συγκεκριμένες εφαρμογές η έξοδος μπορεί να περιλαμβάνει αναγνωριστικά ή δείκτες του προτιμώμενου περιεχομένου, ή ακόμα και μια περίληψη του.
Το TVA-Forum σχήμα για τις προτιμήσεις των χρηστών βασίζεται στο περιγραφικό σχήμα που βρίσκεται ορισμένο στο ISO/IEC 15938-5.
Το σχήμα των προτιμήσεων των χρηστών είναι συσχετισμένο με έναν συγκεκριμένο χρήστη μέσω ενός αναγνωριστικού στοιχείου για τον χρήστη(UserIdentifier). Η βασικότερη οντότητα στο σχήμα, οι Προτιμήσεις Χρήσης(UsagePreferences DS), περιέχουν δύο κύρια

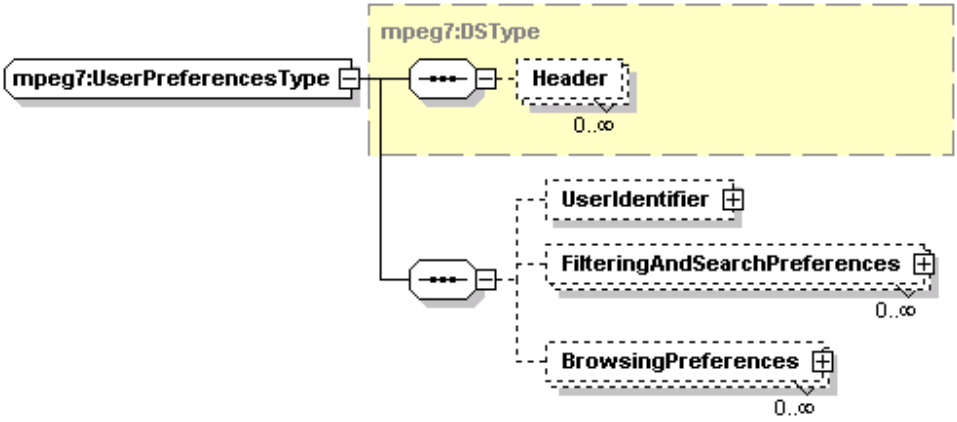
στοιχεία τις Προτιμήσεις Περιήγησης(BrowsingPreferences Ds) και τις Προτιμήσεις Φιλτραρίσματος και Αναζήτησης(FilteringAndSearchPreferenes). Στη παρούσα εργασία ασχοληθήκαμε με τις δεύτερες οι οποίες χρησιμοποιούνται για να ορίσουν τις προτιμήσεις σχετικά με τον τύπο του περιεχομένου προς αναζήτηση, επιλογή και κατανάλωση. Αυτό το σχήμα αποτελείται από τρία επί μέρους σχήματα, το σχήμα των προτιμήσεων κατηγοριοποίησης(ClassificationPreferences DS), το σχήμα των προτιμήσεων δημιουργίας(CreationPreferences DS) και τέλος το σχήμα των προτιμήσεων προέλευσης(SourcePreferences DS).

Το σχήμα των προτιμήσεων χρήσης επιτρέπει στους χρήστες να ορίζουν προτιμήσεις που να ισχύουν για συγκεκριμένο περιεχόμενο, αναφορικά με τον τόπο και τον χρόνο, χρησιμοποιώντας το σχήμα των συνθηκών προτίμησης (PreferenceCondition DS). Επίσης το σχήμα των Προτιμήσεων Χρήσης επιτρέπει στους χρήστες να ορίζουν την σχετική σπουδαιότητα των προτιμήσεων μεταξύ τους. Παρέχεται ακόμη η δυνατότητα στον χρήστη να ορίζει την ιδιωτικότητα ή μη των προτιμήσεών του, καθώς και την δυνατότητα αυτόματης προσαρμογής των προτιμήσεών του από το σύστημα ή όχι. Το σχήμα των Προτιμήσεων Κατηγοριοποίησης χρησιμοποιείται για να ορίζει τις προτιμήσεις που έχουν να κάνουν με την κατηγοριοποίηση του περιεχομένου όπως η γλώσσα, το είδος, η χώρα προέλευσης του περιεχομένου κ.τ.λ. Οι Προτιμήσεις Δημιουργίας χρησιμοποιούνται για να ορίσουν τις προτιμήσεις του χρήστη που σχετίζονται με την δημιουργία του οπτικοακουστικού περιεχομένου όπως ο τίτλος, οι δημιουργοί, ο χρόνος δημιουργίας κ.α. Τέλος οι Προτιμήσεις Προέλευσης χρησιμοποιούνται για να ορίσουν τις προτιμήσεις που αναφέρονται στην προέλευση του περιεχομένου όπως ο εκδότης, το μέσο διάχυσης κ.τ.λ.

Γενικά οι Προτιμήσεις Χρήστη μπορούν να κατασκευαστούν είτε αυτόματα είτε χειροτεχνικά. Στην δεύτερη περίπτωση η περιγραφή στηρίζεται σε ρητή είσοδο από των χρήστη ενώ στην πρώτη οι προτιμήσεις μπορούν να προκύψουν αυτόματα από το ιστορικό Χρήσης του χρήστη.

Η επόμενη ενότητα περιέχει τον ορισμό της σύνταξης και τη σημασιολογία του περιγραφικού σχήματος των Προτιμήσεων Χρήστη.

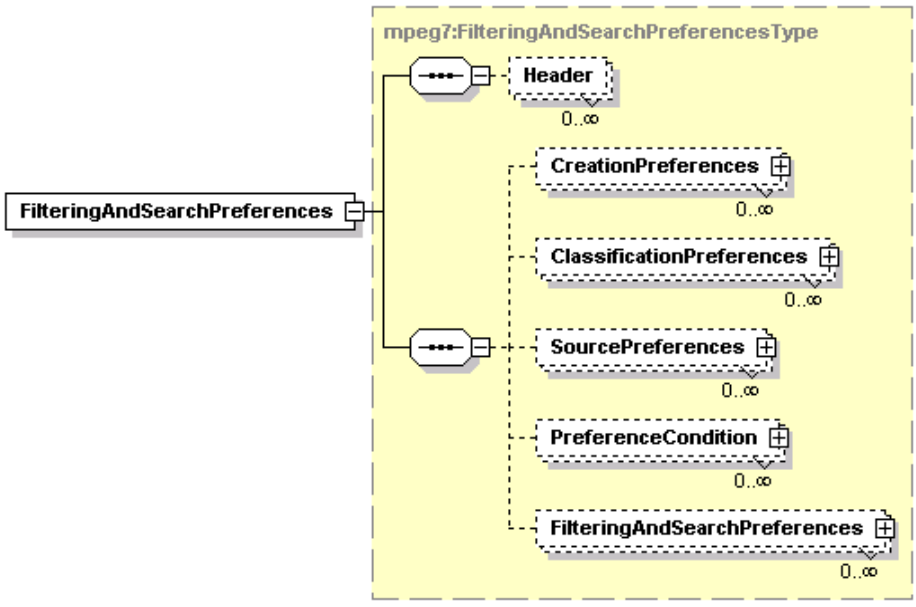
Προτιμήσεις Χρήστη



Σχήμα 10: Σχήμα περιγραφής Προτιμήσεων Χρηστών.

Στο σχήμα φαίνονται τα χαρακτηριστικά που αναφέρθηκαν παραπάνω και σχετίζονται με την κατανάλωση πολυμεσικού υλικού. Οι αντιστοίχιση μεταξύ τις παραπάνω πληροφορίας και την περιγραφή του περιεχομένου είναι αυτή που επιτρέπει την εξατομικευμένη πρόσβαση και κατανάλωση του περιεχομένου.

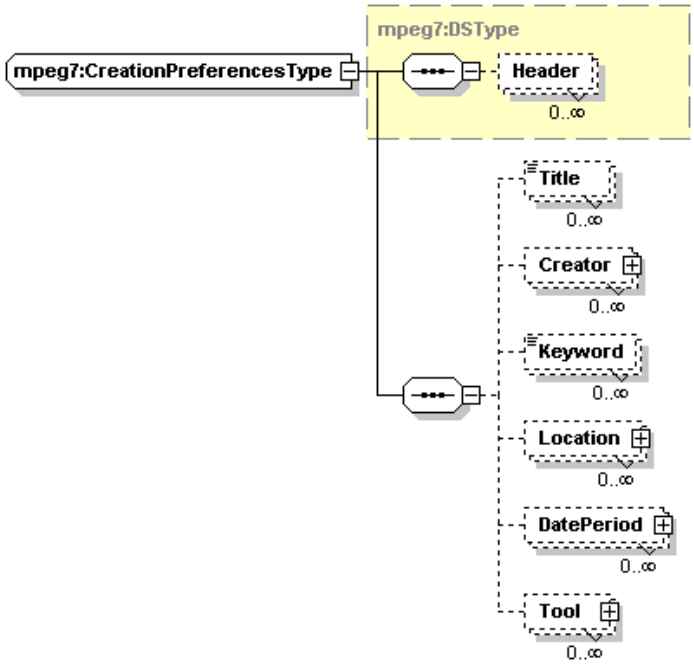
Προτιμήσεις Φιλτραρίσματος και Αναζήτησης(FASP)



Σχήμα 11: Σχήμα περιγραφής Προτιμήσεων Φιλτραρίσματος και Αναζήτησης.

Οι προτιμήσεις Φιλτραρίσματος και αναζήτησης, που στο εξής θα καλούμε FASP ορίζονται για να περιέχουν τις προτιμήσεις του χρήστη σε σχέση με το φιλτράρισμα και την αναζήτηση του οπτικοακουστικού περιεχομένου. Όπως έχουμε ήδη αναφέρει χωρίζονται σε τρεις κατηγορίες. Αξίζει εδώ να υπογραμμιστεί η δυνατότητα που παρέχεται ώστε ένα FASP να μπορεί να περιέχει άλλα FASP σαν παιδιά. Το γεγονός αυτό οδηγεί στην δημιουργία ιεραρχικών δομών από FASP. Κάτι τέτοιο αυξάνει την εκφραστικότητα του σχήματος που έχουμε σαν εργαλείο για την περιγραφή των προτιμήσεων ενός χρήστη. Με αυτόν το τρόπο παρέχεται η δυνατότητα για ορισμό ιεραρχικών σχέσεων μεταξύ των προτιμήσεων. Οι ιεραρχικές δομές αντιπροσωπεύουν εξειδίκευση κατά μήκος της ιεραρχίας από τον πατέρα προς τα φύλλα. Ένα επιπλέον χαρακτηριστικό που περιέχουν τα FASP αντικείμενα είναι αυτό της τιμής προτίμησης(PreferenceValue). Πρόκειται για την τιμή που περιγράφει το σχετικό βάρος των προτιμήσεων μεταξύ τους όταν συνυπάρχουν περισσότερα από ένα FASP ως παιδιά του ίδιου πατέρα. Επίσης εκφράζουν την τιμή προτίμησης για το συνδυασμό των χαρακτηριστικών που περιέχονται κάτω από το FASP (ClassificationPreferences, CreationPreferences, SourcePreferences).

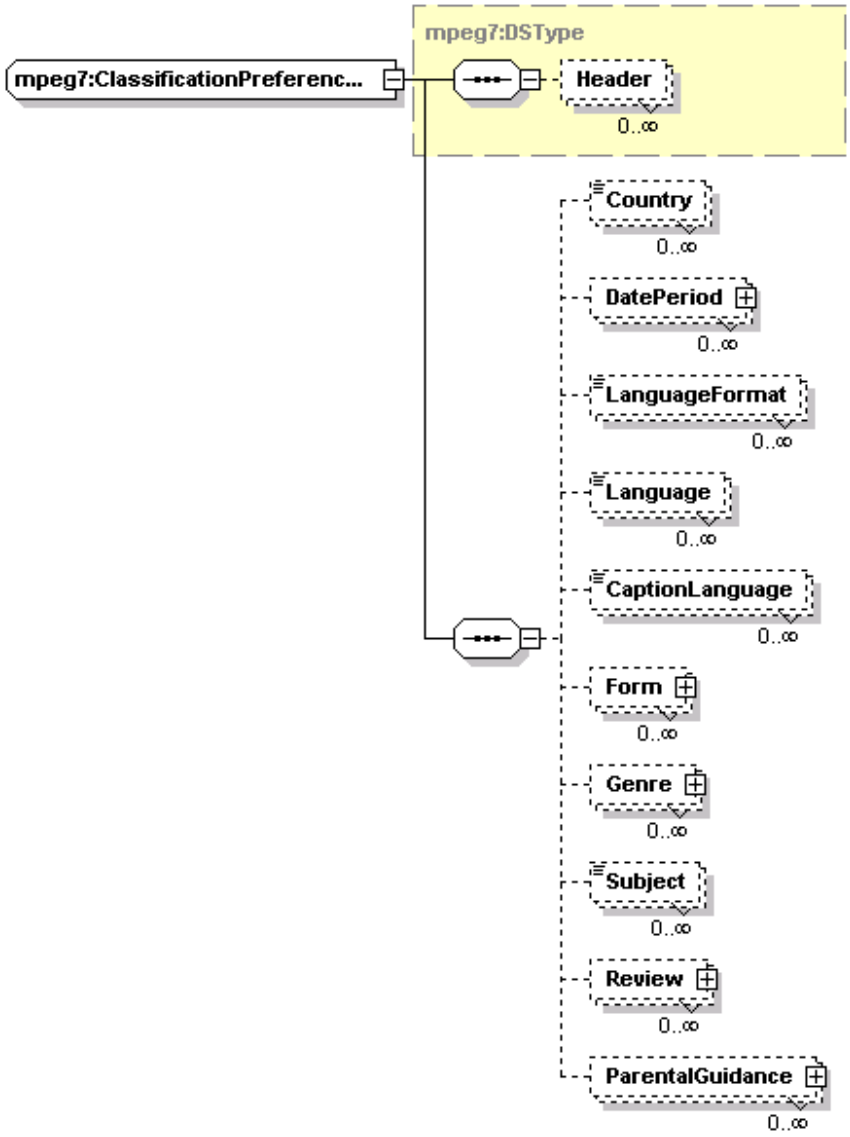
Προτιμήσεις Δημιουργίας



Σχήμα 12: Σχήμα περιγραφής Προτιμήσεων Δημιουργίας.

Πρόκειται για τις προτιμήσεις που αναφέρονται στη σχετική με τη δημιουργία του οπτικοακουστικού περιεχομένου πληροφορία. Τα αντικείμενα που περιέχει φαίνονται στο παραπάνω σχήμα. Επίσης σε κάθε CreationPreferences αντικείμενο υπάρχει η τιμή προτίμησης που υποδηλώνει το σχετικό βάρος όταν περισσότερα του ενός CreationPreferences αντικείμενα συνυπάρχουν σαν αδέλφια.

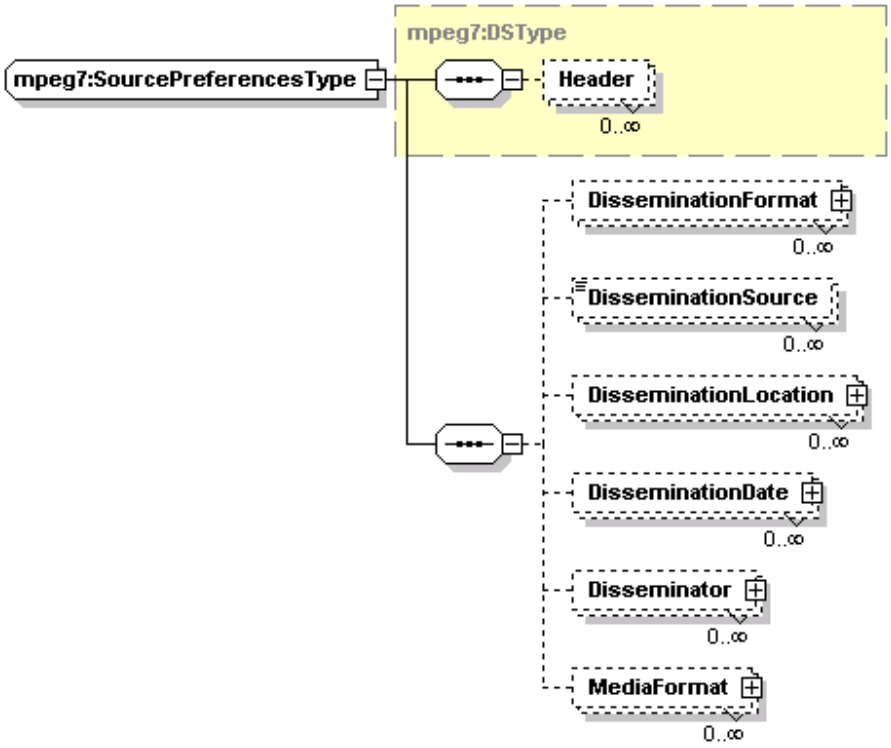
Προτιμήσεις Κατηγοριοποίησης



Σχήμα 13: Σχήμα περιγραφής Προτιμήσεων Κατηγοριοποίησης.

Οι συγκεκριμένες προτιμήσεις όπως φαίνεται αναφέρονται σε όλα εκείνα τα στοιχεία που επιτρέπουν την κατηγοριοποίηση του περιεχομένου. Πάλι περιέχεται μια τιμή προτίμησης για να περιγράφει τη σχετική προτίμηση των ClassificationPreferences μεταξύ τους.

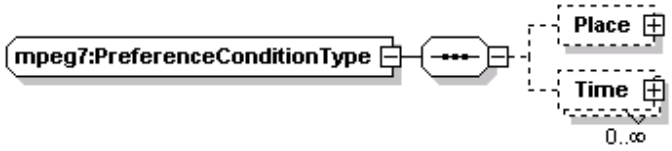
Προτιμήσεις Προέλευσης



Σχήμα 14: Σχήμα περιγραφής Προτιμήσεων Προέλευσης.

Οι Προτιμήσεις Προέλευσης χρησιμοποιούνται για να περιέχουν τις προτιμήσεις που αναφέρονται στην προέλευση του περιεχομένου. Και στα SourcePreferences περιέχεται μια τιμή προτίμησης για να υποδηλώνει το σχετικό βάρος μεταξύ τους.

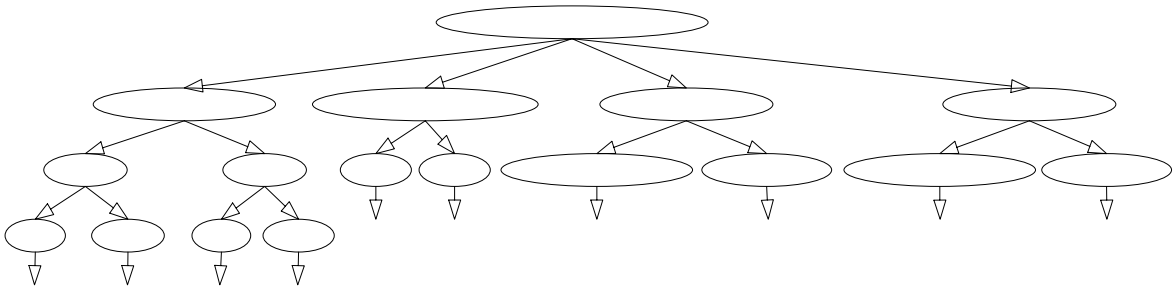
Συνθήκη Προτίμησης



Σχήμα 15: Σχήμα περιγραφής Συνθηκών Προτίμησης.

Αναφέρονται σε ιδιαίτερες συνθήκες που μπορεί να ορίσει ο χρήστης αναφορικά με τον τρόπο και τον χρόνο εφαρμογής των υπόλοιπων προτιμήσεών του.

Για να απλοποιήσουμε την παρουσίαση των παραπάνω θα υιοθετήσουμε την ακόλουθη γραφική αναπαράσταση. Τα τόξα έχουν τις τιμές προτίμησης των αντίστοιχων αντικειμένων:

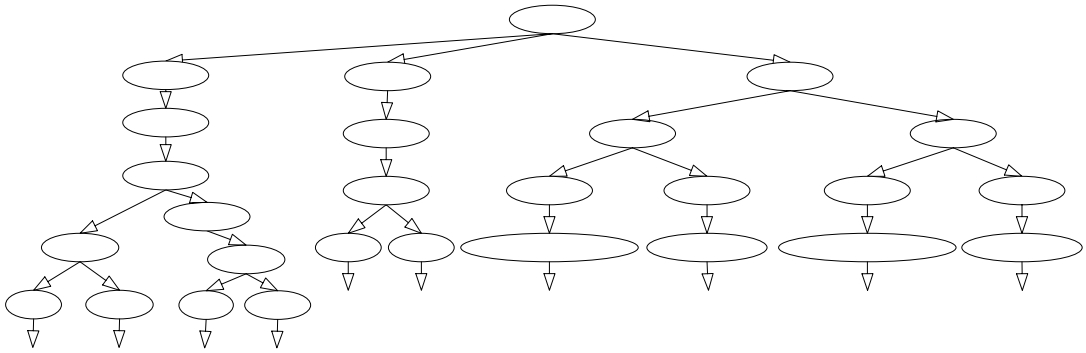


Σχήμα 16: Παράδειγμα δομής Προτιμήσεων Φιλτραρίσματος και αναζήτησης κατά το TVA.

Αλγεβρική Μετάφραση Μοντέλου

Το UP-TV προτείνει κάποια επιπλέον σημασιολογία πάνω στην δομή των FASP προκειμένου να είναι δυνατή η αναπαράσταση τους με αλγεβρικό συντακτικό δέντρο δίτιμης λογικής. Οι αντιστοίχιση γίνεται με τα συνδετικά AND, OR και NOT. Η γενική ιδέα είναι ότι κάθε αντικείμενο αποτελείται από μια λογική σύζευξη των μερών του. Αν ένα μέρος ενός στοιχείου είναι μια λίστα από στοιχεία ίδιου τύπου, τότε αυτό αντιστοιχεί σε λογική διάζευξη των μερών του. Τέλος ένα στοιχείο με αρνητική τιμή προτίμησης αντιστοιχεί σε λογική άρνηση. Στην υλοποίηση που ακολουθήσαμε χρησιμοποιήθηκε το Extended Boolean για μετάφραση της δομής των FASP και δόθηκε ιδιαίτερη προσοχή στις παραμέτρους που ορίζουν την «αυστηρότητα» των AND, OR ώστε να υπάρχει συμφωνία με την επίδραση των διαφόρων χαρακτηριστικών του προφίλ στις προτιμήσεις του χρήστη.

Σύμφωνα με τα παραπάνω οι τιμές προτίμησης είναι σημαντικές μόνο στις περιπτώσεις από λίστες με αντικείμενα ίδιου τύπου. Τότε αντιπροσωπεύουν την σχετική σημαντικότητα του κάθε αντικειμένου της λίστας σε σχέση με τα υπόλοιπα αντικείμενα της λίστας. Το επόμενο δέντρο αντιπροσωπεύει το συντακτικό δέντρο που αντιστοιχεί στο FASP του προηγούμενου σχήματος.



Σχήμα 17: Παράδειγμα αλγεβρικής μετάφρασης του προφίλ από το *σχήμα 16*.

Ένας γενικός μηχανισμός μετάφρασης παρουσιάζεται εδώ. Η αλγεβρική έκφραση που προκύπτει είναι απευθείας εφαρμόσιμη με μοντέλα όπως το Extended Boolean μοντέλο προκειμένου να προκύψει η αντιστοιχία μεταξύ του FASP, του χρήστη και της λίστας των διαθέσιμων προγραμμάτων.

Αναλυτικότερα τα βήματα που ακολουθούνται προχωρώντας από επίπεδο σε επίπεδο είναι τα εξής: πρώτα κάθε FASP κόμβος συσχετίζεται με τη σύζευξη των τεσσάρων δυνατών παιδιών του δηλαδή τις προτιμήσεις δημιουργίας $\{cr_i \mid i=1,\ldots n_{cr}\}$, τις προτιμήσεις κατηγοριοποίησης $\{cl_i \mid i=1,..,n_{cl}\}$, τις προτιμήσεις προέλευσης $\{sr_i \mid i=1,..,n_{sr}\}$ και τέλος άλλους πιθανούς κόμβους-παιδιά FASP $\{fs_i \mid i=1,..,n_{fs}\}$. Η ερώτηση $Q(fs)$ που δημιουργείται είναι η ακόλουθη:

OR

AND

$$Q(fs) = AND\Bigg(\Big(\overset{n_{cr}}{\underset{i=1}{OR}} (Q(cr_i), cr_i.p), 1 \Big), \Big(\overset{n_{cl}}{\underset{i=1}{OR}} (Q(cl_i), cl_i.p), 1 \Big), \Big(\overset{n_{sr}}{\underset{i=1}{OR}} (Q(sr_i), sr_i.p), 1 \Big), \Big(\overset{n_{fs}}{\underset{i=1}{OR}} (Q(fs_i), fs_i.p), 1 \Big) \Bigg)$$

Το αποτέλεσμα είναι μια σύζευξη τεσσάρων διαζευκτικών εκφράσεων που αντιστοιχούν στα αντικείμενα που περιγράψαμε παραπάνω. Στο αμέσως επόμενο επίπεδο χρειάζεται ο ορισμός των $Q(cr)$, $Q(cl)$ και $Q(sr)$ ερωτήσεων που αντιστοιχούν στις προτιμήσεις δημιουργίας, κατηγοριοποίησης και προέλευσης αντίστοιχα.

$$Q(cr) = \underset{f \in \{Title, Creator, Keyword, Location, DatePeriod, Tool\}}{AND} \left(\left(\overset{n_f}{\underset{i=1}{OR}} (Q(f_i), f_i.p), w \right) \right)$$

$$Q(cl) = \underset{f \in \{Country, DatePeriod, LanguageFormat, Language, CaptionLanguage, Form, Genre, Subject, Review, ParentalGuidance\}}{AND} \left(\left(\overset{n_f}{\underset{i=1}{OR}} (Q(f_i), f_i.p), w \right) \right)$$

$$Q(sr) = \underset{f \in \{Dissemination Format, Dissemination Source, Dissemination Location, Dissemination Date, Dissemination or, MediaFormat\}}{AND} \left(\left(\overset{n_f}{\underset{i=1}{OR}} (Q(f_i), f_i.p), w \right) \right)$$

Στις περιπτώσεις αυτές κάθε ερώτηση είναι μια σύζευξη από διαζευκτικές ερωτήσεις. Η σύζευξη γίνεται μεταξύ στοιχείων διαφορετικού τύπου ενώ η διάζευξη μεταξύ στοιχείων ίδιου τύπου. Για τη σύζευξη κατά τη συσχέτιση των διαφορετικών χαρακτηριστικών υποστηρίζεται από το σύστημα χρήση διαφορετικών βαρών για κάθε τύπο χαρακτηριστικού. Παρ' όλα αυτά θα πρέπει να σημειωθεί πως στη παρούσα μορφή του TVA δεν υπάρχει η δυνατότητα για αποθήκευση αυτών των βαρών στη δομή του FASP.

Κεφάλαιο 4: *Περιγραφή Συστήματος*

Στο κεφάλαιο αυτό θα ασχοληθούμε με την περιγραφή της υλοποίησης του συστήματος. Συγκεκριμένα θα περιγραφεί η υλοποίηση κάθε υποσυστήματος με βάση την αρχιτεκτονική που αναλύσαμε στο 2^{ο} κεφάλαιο.

Υποσύστημα Εκτίμησης Προτιμήσεων(UPEM)

Το πρώτο υποσύστημα είναι υπεύθυνο για την αυτόματη εξαγωγή των προτιμήσεων του χρήστη με βάση το ιστορικό των ενεργειών που κατεγράφησαν κατά την αλληλεπίδραση του χρήστη με το σύστημα. Στο σύστημά υπάρχουν σαφώς ορισμένοι όλοι οι τύποι ενεργειών στις οποίες μπορεί να προβεί ένας χρήστης. Το πρώτο βήμα για την εκμετάλλευση αυτής της πληροφορίας είναι το να χωρίσουμε τις ενέργειες με βάση το αν αυτές υποδηλώνουν ικανοποίηση ή όχι για το τμήμα του περιεχομένου που κατανάλωνε ο χρήστης όταν παρατηρήθηκε η αντίστοιχη ενέργεια. Στον επόμενο πίνακα φαίνονται όλες οι ενέργειες, χωρισμένες σε τέσσερις βασικές κατηγορίες όσον αφορά στον τύπο τους και σε δύο κατηγορίες ανάλογα με το αν υποδηλώνουν ικανοποίηση ή όχι του χρήστη για το περιεχόμενο.

Βασική Συμπεριφορά	Ενέργειες με θετική ένδειξη ικανοποίησης	Ενέργειες με αρνητική ένδειξη ικανοποίησης
Επισκόπηση	1.1 PlayRecording 1.2 PlayStream 1.4 Preview 1.5 Pause	1.6 FastForward 1.8 SkipForward 1.10 Mute 1.12

	1.7 Rewind 1.9 SkipBackward 1.11 VolumeUp 1.13 Loop/Repeat 1.14 Shuffle 1.16 SkipToStart 2.1 Zoom 2.2 SlowMotion 2.3 CCOn 2.5 StepBackward 4.2 ScrollUp 4.3 ScrollDown 4.4 ViewGuide 4.7 Search 4.8 SubmitForm 4.9 SubmitQuery	VolumeDown 1.15 SkipToEnd 2.4 StepForward
Αποθήκευση	1.3 Record 1.17 CopyCD 4.5 SavePage 4.6 PrintPage 4.10 Archive 4.11 Select	4.12 Delete
Αναφορά	4.1 ClickThrough	
Σχολιασμός	5.1 Rate 5.2 Annotate/Comment 5.3 Bookmark	

Πίνακας 1: Λίστα όλων των δυνατών ενεργειών ομαδοποιημένων ανά κατηγορία και ένδειξη προτίμησης για τον χρήστη.

Αυτός όμως ο χωρισμός δεν είναι αρκετός. Πρώτα απ'όλα υπάρχει και περαιτέρω ανάγκη για διάκριση των ενεργειών ανάλογα και με τον βαθμό ικανοποίησης που πιθανώς να υποδηλώνουν για τον χρήστη. Στη συνέχεια προκειμένου αυτή η πληροφορία να είναι άμεσα εκμεταλλεύσιμη θα πρέπει να χρησιμοποιηθεί κάποιος πιο αυστηρός τρόπος διάκρισης των ενεργειών. Αυτός δεν είναι άλλος από την απόδοση σε αυτές βαρών που να υποδηλώνουν τον αντίστοιχο βαθμό ικανοποίησης. Για αυτόν το σκοπό χρησιμοποιήθηκαν τιμές από -100 μέχρι 100 τέτοιες ώστε όσο πιο μεγάλη η τιμή τόσο μεγαλύτερος και ο βαθμός ικανοποίησης του χρήστη. Αντίστροφα όσο πιο μικρή αρνητική τιμή τόσο μεγαλύτερη η δυσαρέσκεια του χρήστη από το περιεχόμενο. Τέλος οι μεσαίες τιμές τις κλίμακας υποδηλώνουν ουδέτερη ως αδιάφορη στάση του χρήστη για το περιεχόμενο. Στον επόμενο πίνακα φαίνεται ένα υποσύνολο των ενεργειών που αναφέρονται σε ενέργειες πάνω σε βίντεο. Είναι οι ενέργειες που χρησιμοποιήσαμε και στις οποίες αποδώσαμε ενδεικτικά τα βάρη που φαίνονται στον πίνακα.

ActionType TermID	ActionType Name	ActionType Weight
1.1	PlayRecording	80
1.2	PlayStream	50
1.3	Record	90
1.4	Preview	30
1.5	Pause	50
1.6	FastForward	-20
1.7	Rewind	50
1.8	SkipForward	-30
1.9	SkipBackward	50
1.10	Mute	-10
1.11	VolumeUp	10
1.12	VolumeDown	-10
1.13a	Repeat	60
1.13b	Loop	50
1.14	Shuffle	10
1.15	SkipToEnd	50

1.16	SkipForward	-30
1.17	CopyCd	70
2.1	Zoom	20
2.2	SlowMotion	30
2.4	StepForward	-30

Πίνακας 2: Λίστα ενεργειών με αντιστοιχία βαρών με βάση την προτίμηση που υποδηλώνουν.

Σύμφωνα με το σχήμα του TVA για την Ιστορία Χρήσης, όπως το έχουμε ήδη περιγράψει, υπάρχουν καταγεγραμμένες όλες οι ενέργειες του κάθε χρήστη συσχετισμένες με την πληροφορία για το πρόγραμμα πάνω στο οποίο γίναν και την χρονική στιγμή που συνέβησαν. Με βάση τα βάρη που ορίσαμε παραπάνω, ο πιο άμεσος τρόπος για να συμπεράνουμε την συνολική προτίμηση του χρήστη για ένα πρόγραμμα είναι ομαδοποιώντας όλες τις ενέργειές του ανα πρόγραμμα και παίρνοντας τον μέσο όρο των βαρών τους. Το ίδιο μπορεί να εφαρμοστεί και στην περίπτωση που υπάρχει πληροφορία για κατάτμηση(segmentation information) των προγραμμάτων οπότε και μπορούμε να ομαδοποιήσουμε αντίστοιχα όλες του ενέργειες του κάθε τμήματος των προγραμμάτων. Από μαθηματικής άποψης ο τύπος που ουσιαστικά εφαρμόζουμε είναι ο :

$$p_u(c) = \frac{\sum_{i=1}^{n_{u,c}} w(a_i.type)}{n_{u,c}}$$

όπου $p_u(c)$ είναι η εκτίμηση της προτίμησης του χρήστη *u* για το πρόγραμμα/τμήμα *c*, a_i *είναι η* i-οστη ενέργεια του χρήστη u πάνω στο πρόγραμμα/τμήμα *c*, $a_i.type$ είναι ο τύπος της ενέργειας a_i, , $w(a_i.type)$ είναι το βάρος που έχει αντιστοιχηθεί στον αντίστοιχο τύπο ενέργειας και $n_{u,c}$ το πλήθος των ενεργειών που κατεγράφησαν κατά τη διάρκειά του. Επιπλέον αυτού του μηχανισμού έχουν αναπτυχθεί αλγόριθμοι που εντοπίζουν το ακριβές σημείο μέσα σε κάθε πρόγραμμα ή τμήμα προγράμματος που παρατηρήθηκαν συγκεκριμένες ενέργειες. Επικεντρώνοντας στις ενέργειες που υποδηλώνουν πραγματική κατανάλωση του περιεχομένου, όπως *play, record* και αξιοποιώντας τις ενέργειες που υποδηλώνουν περιήγηση πάνω στο περιεχόμενο, όπως *rewind, fastforward*, ώστε να γνωρίζουμε ανα πάσα στιγμή σε πιο σημείου στη ροή του βίντεο βρίσκεται ο χρήστης,

μπορούμε να υπολογίσουμε μια σειρά από επιπλέον ενδιαφέρουσες πληροφορίες. Με αυτόν τον τρόπο λοιπόν μπορούμε να υπολογίσουμε :

- το πραγματικό ποσοστό κάθε προγράμματος ή τμήματος προγράμματος που είδε ο χρήστης
- τον αριθμό των φορών που είδε ο χρήστης κάθε τμήμα κάθε προγράμματος ή τμήματος προγράμματος.

Μετά, αυτήν την πληροφορία μπορούμε να την ενσωματώσουμε στην τελική τιμή προτίμησης που υπολογίζεται για κάθε πρόγραμμα. Τέλος μπορούμε να την συσχετίσουμε με τυχόν υπάρχουσα πληροφορία για κατάτμηση και με συγκεκριμένα πλέον χαρακτηριστικά των προγραμμάτων. Συγκεκριμένα για κάθε τμήμα προγράμματος είναι δυνατό να υπάρχουν διαθέσιμα μεταδεδομένα, όπως λέξεις κλειδιά, τα οποία μπορούν να χρησιμοποιηθούν κατά την κατασκευή του προφίλ.

Υποσύστημα Κατασκευής και Προσαρμογής Προφίλ(FCAM)

Στην προηγούμενη ενότητα περιγράψαμε τη μεθοδολογία εκτίμησης των προτιμήσεων των χρηστών για μια ομάδα προγραμμάτων με βάση την Ιστορία Χρήσης που έχει καταγραφεί για κάθε χρήστη. Το αποτέλεσμα μιας τέτοιας διαδικασίας είναι μια λίστα προγραμμάτων μαζί με κάποιο βαθμό προτίμησης , που αντιπροσωπεύει την σχετική προτίμηση του χρήστη για το εκάστοτε πρόγραμμα. Μια ανάλογη λίστα με τις προτιμήσεις των χρηστών θα μπορούσε να έχει προκύψει ύστερα από ρητή καταγραφή των προτιμήσεων των χρηστών. Και στις δύο περιπτώσεις, οι προτιμήσεις των χρηστών αποτελούν πληροφορία πολύτιμη για την περιγραφή του χρήστη σε όρους που θα επιτρέψουν την εξατομίκευση του συστήματος και θα παράσχουν την δυνατότητα για αποδοτική αναζήτηση και φιλτράρισμα του διαθέσιμου περιεχομένου. Για να είναι αυτό εφικτό, θα πρέπει οι προτιμήσεις του χρήστη να καταγραφούν με κάποιον επίσημο τρόπο. Όπως ήδη έχουμε πριγράψει(κεφ. *), το εργαλείο που υπάρχει για αυτόν το σκοπό είναι το σχήμα μεταδεδομένων των Προτιμήσεων των Χρηστών.

Σε ότι ακολουθεί θα περιγράψουμε τους μηχανισμούς που έχουν αναπτυχθεί τόσο για την κατασκευή του προφίλ του χρήστη όσο και για την αυτόματη αναπροσαρμογή του όταν προκύπτουν νέα δεδομένα για τον χρήστη.

Σύμφωνα με τα εργαλεία που παρέχονται από το TVA ένα τέτοιο προφίλ βασίζεται στην χρήση των αντικειμένων τύπου FASP(Filtering And Search Preferences) και στις δυνατότητες και τα χαρακτηριστικά που αυτά παρέχουν. Αξίζει να σημειωθεί πως ενώ υπάρχει ακριβής περιγραφή των FASP και των χαρακτηριστικών τους τόσο από δομική όσο και από σημασιολογική άποψη, η χρήση τους ως δομικά στοιχεία για την κατασκευή του προφίλ ενός χρήστη δίνει μια ελευθερία δυνατοτήτων και επιλογών για να αξιοποιηθεί είτε με τη δημιουργία διαφορετικών δομών είτε με την διαφορετική αλγεβρική ερμηνεία της πληροφορίας που ενσωματώνουν κάθε φορά. Στο παρόν θα ασχοληθούμε με τους διαφορετικούς τύπους δομών που αναπτύχθηκαν και υποστηρίζονται από το σύστημα για την κατασκευή του προφίλ ενός χρήστη.

Η δομή του προφίλ των χρηστών μπορεί να είναι είτε επίπεδη είτε ιεραρχική. Στην πρώτη περίπτωση η δομή περιλαμβάνει έναν FASP κόμβο χωρίς FASP 'παιδιά'. Στην δεύτερη περίπτωση, των ιεραρχικών FASP, είναι δυνατή η δημιουργία ενός δέντρου από FASP κόμβους. Απαραίτητη προϋπόθεση είναι ο καθορισμός των κριτηρίων βάση των οποίων θα κτισθεί η ιεραρχία. Στη συνέχεια θα αναφερθούμε αναλυτικά στη μεθοδολογία και τους μηχανισμούς που έχουν αναπτυχθεί προκειμένου να υποστηριχτούν τα παραπάνω μοντέλα προφίλ.

Προφίλ Επίπεδης Δομής

Πρόκειται για την απλούστερη μορφή που μπορεί να έχει το προφίλ ενός χρήστη και όπως ήδη έχουμε αναφέρει αποτελείται από ένα αντικείμενο τύπου FASP το οποίο μπορεί να περιλαμβάνει έναν οποιοδήποτε συνδυασμό από προτιμήσεις Δημιουργίας, Κατηγοριοποίησης και Προέλευσης, προτιμήσεις που αποθηκεύονται στα αντίστοιχα

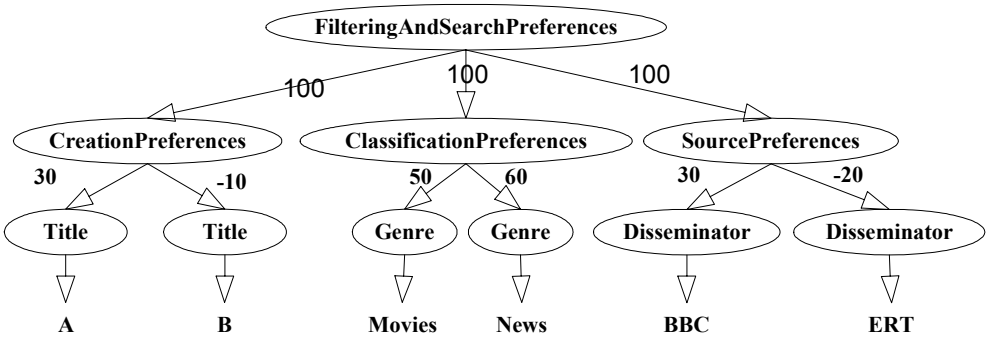
αντικείμενα τύπου *ClassificationPreferences*, *CreationPreferences* και *SourcePreferences*. Η ολοκληρωμένη διαδικασία αποτελείται από τα εξής δύο βήματα:

1. Τη δημιουργία ενός νέου FASP με βάση τις προτιμήσεις για περιεχόμενο ενός συγκεκριμένου χρήστη.
2. Την αφομοίωση νέων προτιμήσεων από το ήδη υπάρχων προφίλ του χρήστη. Συγκεκριμένα αναφερόμαστε στο συνδυασμό τυχόν υπάρχοντος FASP με ένα καινούριο που έχει προκύψει από πιο πρόσφατες προτιμήσεις του χρήστη.

Δημιουργία Προφίλ Επίπεδης Δομής

Το πρώτο βήμα της διαδικασίας είναι η εύρεση όλων των μεταδεδομένων των προγραμμάτων για τα οποία υπάρχει αντίστοιχη τιμή προτίμησης από τον χρήστη. Με αυτό τον τρόπο συγκεντρώνονται όλα τα χαρακτηριστικά των προγραμμάτων(κατηγορία, ηθοποιοί, γλώσσα κ.τ.λ.) που έχει αξιολογήσει ο χρήστης. Για κάθε τιμή ενός χαρακτηριστικού θα πρέπει να υπάρξει μια εκτίμηση της προτίμησης που έχει ο χρήστης για αυτό. Για αυτό το σκοπό, το επόμενο βήμα είναι ο υπολογισμός ενός σχετικού βάρους για κάθε τιμή όλως των χαρακτηριστικών των προγραμμάτων. Ακολουθεί λοιπόν ο υπολογισμός μιας τιμής προτίμησης για κάθε ηθοποιό για παράδειγμα. Και η διαδικασία επαναλαμβάνεται για όλες τις τιμές όλων των χαρακτηριστικών που εμφανίζονται στα προγράμματα που έχει δει ο χρήστης. Οι τελικές τιμές αποθηκεύονται σαν τιμές προτίμησης για τα αντίστοιχα χαρακτηριστικά στο προφίλ του χρήστη. Ο τρόπος εκτίμησης της προτίμησης του χρήστη για κάθε χαρακτηριστικό μπορεί να γίνει με πολλούς εναλλακτικούς τρόπους. Ενδεικτικά αναφέρουμε πως απλούστερος τρόπος θα ήταν η απόδοση σε κάθε χαρακτηριστικό της μέσης τιμής των προτιμήσεων των προγραμμάτων στα οποία εμφανίστηκε. Η αναζήτηση πιο σύνθετων μεθοδολογιών με στόχο την καλύτερη απόδοση του συστήματος είναι αντικείμενο έρευνας και στα πλαίσια της διπλωματικής προσεγγίστηκε με την διεξαγωγή σχετικών πειραμάτων η παρουσίαση των οποίων γίνεται σε επόμενο κεφάλαιο. Εδώ θα θεωρήσουμε πως ανεξαρτήτως της μεθόδου υπολογισμού των τιμών προτίμησης των χαρακτηριστικών, αυτές είναι δεδομένες και θα συνεχίσουμε με τα υπόλοιπα βήματα της κατασκευής του προφίλ ενός χρήστη.

Αφού έχει γίνει ο υπολογισμός των προτιμήσεων για τα διάφορα χαρακτηριστικά και για την δημιουργία του προφίλ, δημιουργείται ένας καινούριος κόμβος FASP και κάτω από αυτόν προστίθενται ανάλογα με τον τύπο τους όλα τα χαρακτηριστικά μαζί με το βαθμό προτίμησής τους και ανάλογα με το αν αυτά χαρακτηρίζουν την δημιουργία, την κατηγορία, η την προέλευση του περιεχομένου. Ένα παράδειγμα μιας τέτοιας δομής φαίνεται στο αμέσως επόμενο σχήμα:



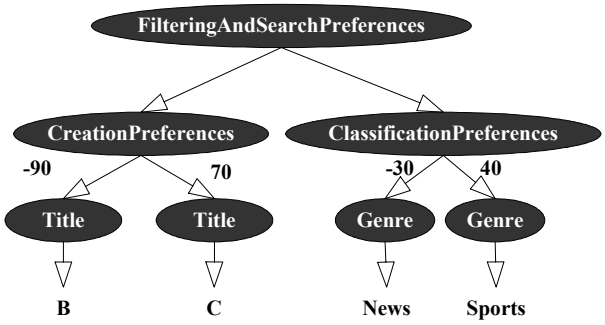
Σχήμα 18: Παράδειγμα Προτιμήσεων Φιλτραρίσματος και Αναζήτησης.

Στα φύλλα, βρίσκονται οι τιμές όλων των χαρακτηριστικών που εμφανίστηκαν στην ομάδα προγραμμάτων για την οποία διαθέτουμε τις προτιμήσεις του χρήστη, ενώ οι τιμές πάνω στα βέλη αντιπροσωπεύουν τις τιμές προτίμησης για τα αντίστοιχα χαρακτηριστικά. Η κατασκευή ενός τέτοιου προφίλ έχει υλοποιηθεί σε MySQL και όλη η δομή αποθηκεύεται στην TVA συμβατή βάση που έχει δημιουργηθεί για το πρόγραμμα up-TV.

Αναπροσαρμογή Προφίλ Επίπεδης Δομής

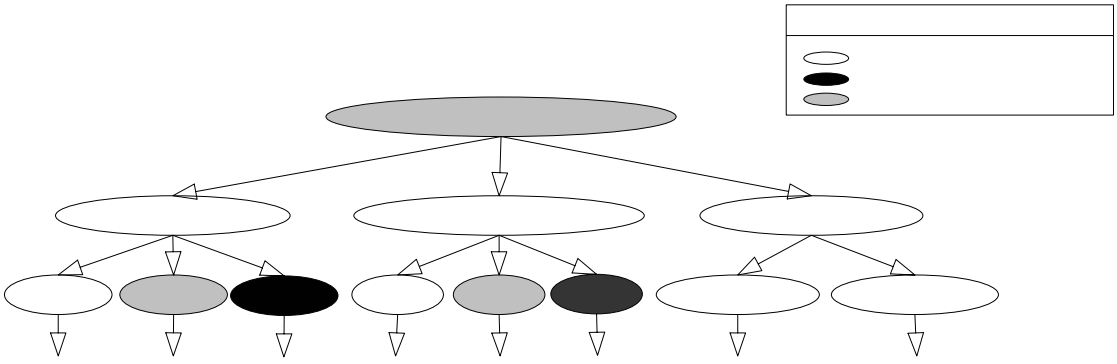
Όταν δεν είναι η πρώτη φορά δημιουργίας προφίλ για το χρήστη, τότε ένα προφίλ επίπεδης δομής είναι δυνατό να υπάρχει ήδη αποθηκευμένο για τον χρήστη. Σε μια τέτοια περίπτωση το σύστημα πρέπει να έχει την δυνατότητα της προσαρμογής των παλιών προτιμήσεων του χρήστη λαμβάνοντας υπόψη της καινούριες. Η διαδικασία στοχεύει ουσιαστικά στην προσαρμογή των τιμών προτίμησης που υπήρχαν στα χαρακτηριστικά του παλιού προφίλ συνδυάζοντάς τες με τις τιμές προτίμησης που υπολογίστηκαν στο καινούριο προφίλ του χρήστη. Τα βήματα που ακολουθούνται περιλαμβάνουν πρώτα τον έλεγχο για την ύπαρξη επίπεδου προφίλ για τον χρήστη, στη συνέχεια την αναζήτηση ανάμεσα στα χαρακτηριστικά που υπάρχουν ήδη στο προφίλ του χρήστη, για αντιστοιχίες με τα

καινούρια που έχουν προκύψει. Ο συνδυασμός μιας παλιάς τιμής προτίμησης για ένα χαρακτηριστικό, με μια καινούρια γίνεται με βάση κάποιον παράγοντα που ας ονομάζουμε *σ*. Τότε για κάθε χαρακτηριστικό η προσαρμοσμένη τιμή προτίμησης π θα είναι $\pi = \sigma \pi_o + (1-\sigma) \pi_v$. Όπου π_o και π_v οι παλιά και η νέα τιμή προτίμησης αντίστοιχα. Για να δούμε και σχηματικά τη διαδικασία ας θεωρήσουμε το FASP του παρακάτω σχήματος σαν το καινούριο για τον χρήστη, του οποίου το παλιό FASP είναι αυτό του προηγούμενου σχήματος:



Σχήμα 19: Νέο FASP που θα χρησιμοποιηθεί για την αναπροσαρμογή αυτού από το *σχήμα 18*.

Παρατηρούμε πως τα δύο FASP έχουν ως κοινά χαρακτηριστικά των τίτλο B και την κατηγορία News. Για αυτά τα κοινά στοιχεία θα γίνει συνδυασμός των τιμών προτίμησής τους, ενώ για αυτά που δεν υπάρχει αντιστοιχία, τα μεν παλιά θα μείνουν ως έχουν τα δε καινούρια θα προστεθούν ως έχουν στο παλιό FASP. Το αποτέλεσμα που θα προκύψει θα είναι αυτό που φαίνεται στο επόμενο σχήμα:



Σχήμα 20: FASP που προκύπτει από το συνδυασμό αυτών του *σχήματος 18* και *19*.

Ένα προφίλ σαν το παραπάνω είναι αρκετά γενικό και οδηγεί σε ένα όλο και μεγαλύτερο σύνολο προγραμμάτων όσο μεγαλώνει ο αριθμός των στοιχείων που περιλαμβάνει. Επιπλέον μεταφράζεται σε διάζευξη όλων των χαρακτηριστικών ίδιου τύπου και αυτή είναι και η επικρατούσα σχέση σε αυτόν τον τύπο FASP. Παρόλα αυτά, πολλές φορές υπάρχει η ανάγκη για αποτύπωση συσχετίσεων μεταξύ χαρακτηριστικών και την δημιουργία φίλτρων που θα περιορίζουν πιο αποδοτικά το σύνολο των προγραμμάτων που επιστρέφεται στον χρήστη. Κυρίως για τους παραπάνω λόγους έχει αναπτυχθεί μεθοδολογία για την δημιουργία ιεραρχικών προφίλ, που εκμεταλλεύονται την εξειδίκευση μεταξύ των διάφορων χαρακτηριστικών. Την μεθοδολογία αυτή αναλύουμε στην αμέσως επόμενη ενότητα.

Προφίλ Ιεραρχικής Δομής

Από τον ορισμό των FASP δίνεται η δυνατότητα δημιουργίας ιεραρχιών από FASP. Μια τέτοια ιεραρχία κατασκευάζεται για να δηλώσει είτε τη σχέση συνόλου και υποσυνόλου μεταξύ γονέα και παιδιού είτε ακόμα και τη σχέση εξειδίκευσης πηγαίνοντας από την ρίζα προς τα φύλλα. Ας μην ξεχνάμε, σε αυτό το σημείο, την σύζευξη που υπάρχει μεταξύ ενός FASP και των 'παιδιών' του. Με αυτόν τον τρόπο είναι εφικτή η ιεραρχική δόμηση των προτιμήσεων του χρήστη αλλά και η επέκταση της δυνατότητας για αποτύπωση συσχετίσεων μεταξύ διαφορετικών χαρακτηριστικών αλλά και προτύπων τέτοιων συσχετίσεων. Έτσι υπάρχει μια πληθώρα δυνατοτήτων όταν καλούμαστε να επιλέξουμε τον τρόπο με τον οποίο θα δημιουργηθεί μια ιεραρχική δομή των προτιμήσεων ενός χρήστη. Το κυρίως ζητούμενο είναι το κριτήριο το οποίο θα επιβάλει την δομή και κατ' επέκταση την ιεραρχική σχέση μεταξύ των διαφόρων κόμβων του υπό κατασκευή δέντρου. Με λίγα λόγια, κάθε προσπάθεια για δημιουργία ιεραρχικού προφίλ θα πρέπει να υπακούει και σε κάποια σαφώς ορισμένη σημασιολογία. Μια πρώτη αναζήτηση οδηγεί σύντομα στην υιοθέτηση οποιουδήποτε χαρακτηριστικού διαθέτει από την φύση του το χαρακτηριστικό της ιεραρχίας. Σαν τέτοιο στην περίπτωση των προγραμμάτων της ψηφιακής τηλεόρασης παρουσιάζεται μόνο το χαρακτηριστικό της κατηγορίας(Genre) στην οποία ανήκει κάθε πρόγραμμα. Αυτό είναι ίσως και το σημαντικότερο χαρακτηριστικό στην προσπάθεια περιγραφής των προτιμήσεων του χρήστη σε περιβάλλον ψηφιακής τηλεόρασης. Θα

ξεκινήσουμε λοιπόν με την περιγραφή της δημιουργίας ιεραρχικού προφίλ με βάση την κατηγορία των προγραμμάτων και θα συνεχίσουμε με την περιγραφή πιο σύνθετων περιπτώσεων δημιουργίας ιεραρχικών προφίλ βασισμένων σε κάποιο πρότυπο ιεραρχίας.

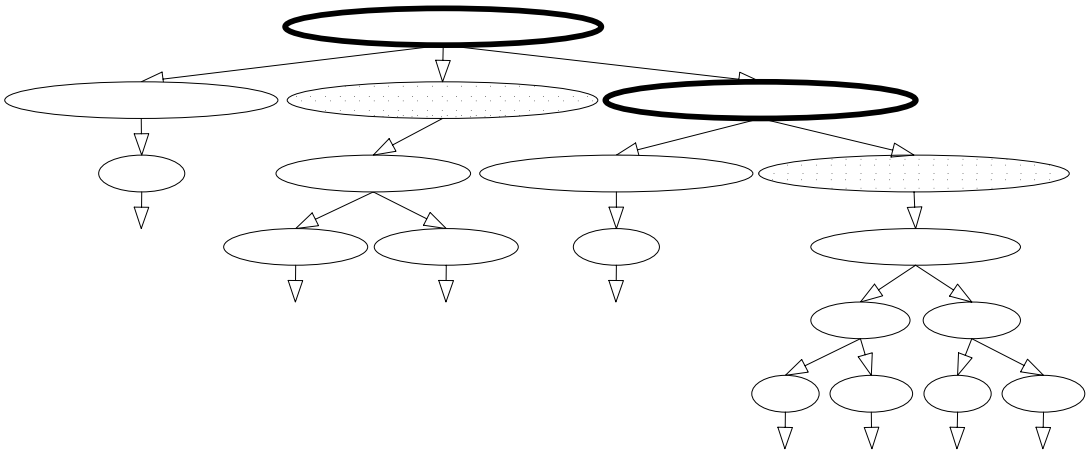
Δημιουργία Προφίλ με βάση την Κατηγορία των Προγραμμάτων

Σύμφωνα με το σχήμα ταξινόμησης του TVA, οι κατηγορίες των προγραμμάτων είναι χωρισμένες και ομαδοποιημένες θεματικά ενώ ακολουθούν και μια ιεραρχική οργάνωση από πιο γενικές κατηγορίες σε πιο ειδικές. Όταν λοιπόν θέλουμε να δημιουργήσουμε μια ιεραρχία από FASP ακολουθώντας αυτή των κατηγοριών των προγραμμάτων η δομή των FASP θα είναι ακριβώς αντίστοιχη με αυτή των κατηγοριών. Για παράδειγμα, οι *κωμωδίες(comedy)* είναι υποκατηγορία των *ταινιών(movie)*. Ακριβώς αντίστοιχα θα περιμένουμε να δούμε ένα FASP που αντιστοιχεί στις *ταινίες* να έχει σαν 'παιδί' FASP κάποιο που να αντιστοιχεί στις *κωμωδίες*. Φυσικά, οι κατηγορίες των προγραμμάτων που μας ενδιαφέρουν είναι αυτές που αντιπροσωπεύουν τα προγράμματα που έχει καταναλώσει ο χρήστης. Ας δούμε όμως από την αρχή τα βήματα που ακολουθούνται προκειμένου να προκύψει κάποιο ιεραρχικό προφίλ για τον χρήστη.

1) Ας θυμηθούμε πρώτα απ' όλα πως αυτό που υπάρχει αρχικά είναι μια λίστα με όλα τα προγράμματα για τα οποία υπάρχει η προτίμηση του χρήστη. Όλα τα προγράμματα ομαδοποιούνται ανάλογα με την/ις κατηγορία/ες στην οποία ανήκουν. Στη συνέχεια η διαδικασία που ακολουθείται είναι ανάλογη με αυτήν που εφαρμόστηκε στα επίπεδα FASP, μόνο που τώρα εφαρμόζεται για κάθε κατηγορία επαναληπτικά. Πιο συγκεκριμένα, οι τιμές προτίμησης κάθε χαρακτηριστικού στα διάφορα προγράμματα, που ανήκουν σε κάθε μια κατηγορία, υπολογίζονται με τον ίδιο τρόπο που ακολουθείται και στα προφίλ επίπεδης δομής μόνο που τώρα στηρίζεται στα προγράμματα που ανήκουν στην εκάστοτε κατηγορία προγραμμάτων. Η διαδικασία επαναλαμβάνεται για όλα τα χαρακτηριστικά και για όλες τις κατηγορίες των προγραμμάτων. Με αυτόν τον τρόπο καταλήγουμε σε μια τιμή προτίμησης για κάθε χαρακτηριστικό και για κάθε κατηγορία στην οποία μπορεί να εμφανίστηκε.
2) Στη συνέχεια δημιουργείται μια ιεραρχία από FASP η οποία αποτελείται από κόμβους δύο τύπων: αυτούς που αποτελούν τον κορμό της ιεραρχίας και περιέχουν μόνο ένα Classification Preference με την κατηγορία του εκάστοτε κόμβου και άλλα

FASP παιδιά. Ένα από αυτά τα παιδιά είναι ο δεύτερος τύπου FASP που υπάρχει για να φέρει όλη την πληροφορία για τις τιμές προτίμησης των χαρακτηριστικών των προγραμμάτων που ανήκουν στην κατηγορία του πατέρα.

Στο επόμενο σχήμα υπάρχει μια ενδεικτική ιεραρχία για έναν χρήστη:



Σχήμα 21: Παράδειγμα Ιεραρχικού με βάση το Genre προφίλ.

Όπως μπορεί κανείς να παρατηρήσει οι κύριοι κόμβοι της ιεραρχίας(με την έντονη γραμμή) περιέχουνε ένα ClassificationPreferences το οποίο με την σειρά του περιέχει το στοιχείο με την κατηγορία που αντιστοιχεί στο FASP(Movies, Comedy). Επίσης παρατηρούμε την μεταξύ τους σχέση που είναι πατέρα-παιδιού. Τέλος κάτω από τους κύριους κόμβους υπάρχει ένας κόμβος FASP(οι γκρι κόμβοι) ο οποίος περιέχει όλα τα χαρακτηριστικά των προγραμμάτων για τα οποία είχε υπολογιστεί κάποιος βαθμός προτίμησης.

Αναπροσαρμογή Ιεραρχικού Προφίλ

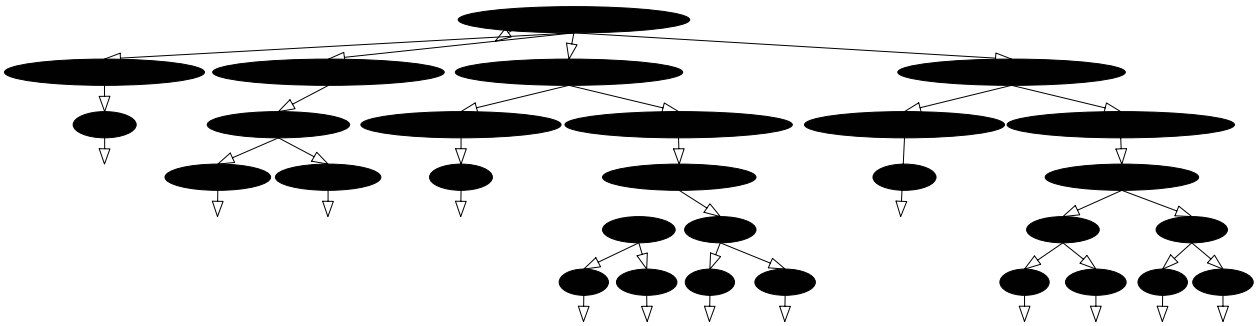
Όταν ο χρήστης έχει ένα ιεραρχικό με βάση την κατηγορία των προγραμμάτων προφίλ, υπάρχει δυνατότητα της αναπροσαρμογής του κάθε φορά που προκύπτουν νέες προτιμήσεις για τον χρήστη. Με στόχο λοιπόν αυτόν, το πρώτο βήμα μιας τέτοιας διαδικασίας είναι να αποθηκευτούν και οι νέες προτιμήσεις σε μια αντίστοιχη δομή. Το επόμενο και σημαντικότερο βήμα της διαδικασίας ας το δούμε μέσα από ένα παράδειγμα. Συγκεκριμένα αν θεωρήσουμε σαν το παλιό FASP του χρήστη αυτό του προηγούμενου

ClassificationPreferences

Genre

Movies

σχήματος και σαν το καινούριο το αμέσως παρακάτω τα βήματα που θα ακολουθηθούν θα είναι:



Σχήμα 22: Νέο Ιεραρχικό FASP που θα χρησιμοποιηθεί για το συνδυασμό με αυτό του *σχήματος 21.*

1) έλεγχος μέσα στο προφίλ ενός χρήστη για εύρεση μιας ιεραρχίας ορθά δομημένης με βάση την κατηγορία των προγραμμάτων[1].

2) Έλεγχος της νέας ιεραρχίας και εύρεση αντιστοιχίας κόμβων του νέου δέντρου μέσα στο παλιό. Όπως ο κόμβος *Movies* και *Comedy* στην περίπτωσή μας. Στην συνέχεια γίνεται εύρεση των κοινών χαρακτηριστικών από αυτά που περιέχουν. Για αυτά που είναι κοινά[2] γίνεται προσαρμογή της προτίμηση με κάποιο βάρος, όπως και στα επίπεδα FASP, ενώ για κάποια νέα που δεν υπήρχαν[3] γίνεται προσθήκη τους στον κόμβο που ανήκουν. Το ίδιο συμβαίνει και στην περίπτωση που υπάρχει ολόκληρος κόμβος FASP που δεν υπήρχε στη προηγούμενη ιεραρχία οπότε και πρέπει να κρεμαστεί από τον σωστό πατέρα.

Συνοπτικά αυτό που γίνεται είναι οι αντιστοίχιση όμοιων κόμβων και η αναπροσαρμογή των κοινών χαρακτηριστικών. Το αποτέλεσμα που προκύπτει σύμφωνα με τα παραπάνω φαίνεται στο αμέσως επόμενο σχήμα:

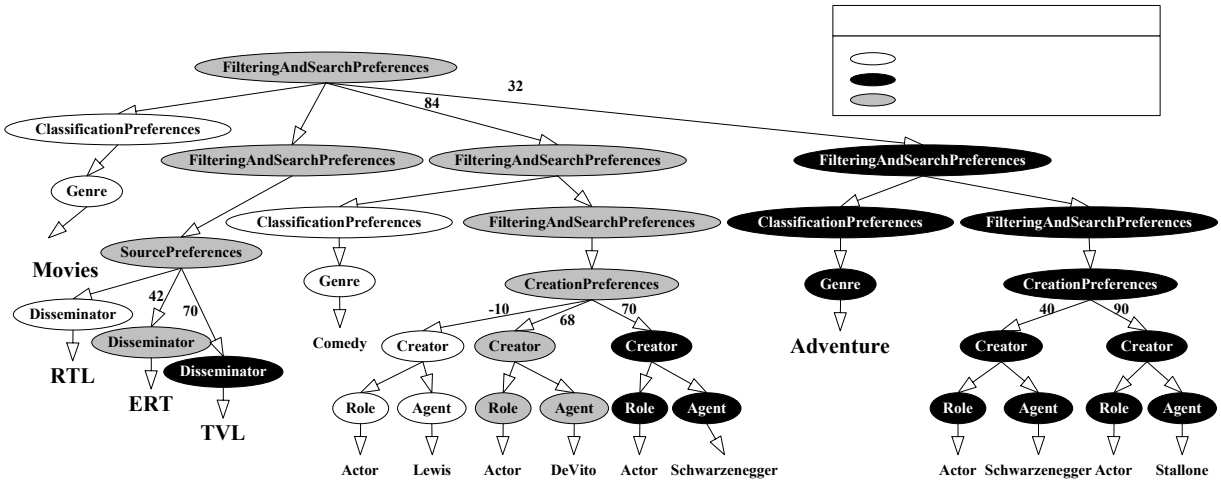
20 **70**

Movies

[1] Δεδομένου ότι το προφίλ ενός χρήστη μπορεί να περιέχει πολλά FASP ρίζες και όχι απαραίτητα ίδιου τύπου όσον αφορά την δομή τους.

[2] Εδώ κοινό είναι π.χ. η ERT **ERT** **TVL**

[3] όπως η TVL.



Σχήμα 23: Αναπροσαρμοσμένο FASP που προκύπτει από το συνδυασμό αυτών των *σχημάτων 21,22.*

Για τις λεπτομέρειες της υλοποίησης αναφέρουμε πως η όλη διαδικασία γίνεται σε java ενώ τα προφίλ βρίσκονται αποθηκευμένα στη βάση δεδομένων σύμφωνα με το TVA του up-TV. Η διαδικασία γίνεται στη μνήμη με χρήση των δομών του Breeze ώστε να υπάρχει πάντα η δυνατότητα αποθήκευσης στη βάση ή σε TVA-σύμφωνο XML έγγραφο του νέου, του παλιού ή του προσαρμοσμένου προφίλ του χρήστη.

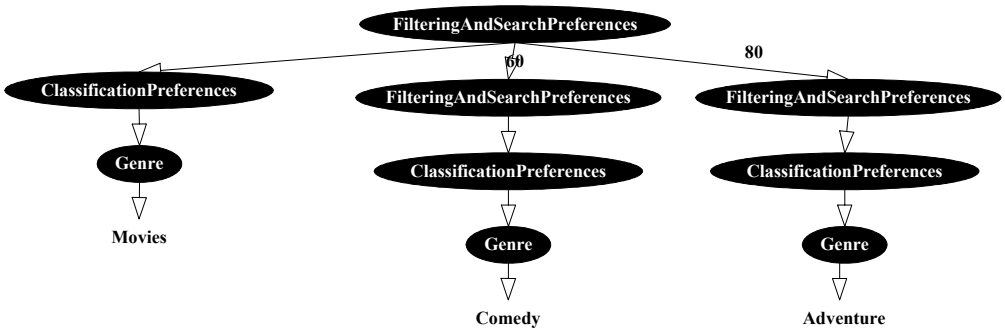
Προφίλ Ιεραρχικής Δομής Βάση Προτύπου

Η πληθώρα των χαρακτηριστικών αλλά και των στοιχείων που χαρακτηρίζουν κάθε πρόγραμμα, και μάλιστα τόσο αυτών που έχουν να κάνουν με την κατηγοριοποίηση του προγράμματος αλλά και εξίσου σημαντικών που αφορούν στην δημιουργία αλλά και την προέλευση του προγράμματος, μας δίνει το έναυσμα για να αναζητήσουμε και να ερευνήσουμε τις δυνατότητες για δόμηση ιεραρχιών με βάση άλλων στοιχείων ως κριτήρια για τη δημιουργία της ιεραρχίας. Εδώ θα πρέπει να διευκρινιστεί πως η επιλογή αυτών των στοιχείων μπορεί να γίνει με πολλούς τρόπους τόσο στην απλή μορφή της επιλογής προσωπικών κριτηρίων από τον ίδιο τον χρήστη[4], όσο και με χρήση προχωρημένων

[4] Εφόσον υπάρχει το interface που να του δίνει τέτοια δυνατότητα.

τεχνικών που στόχο μπορεί να έχουν την αποτύπωση συνδέσεων(associations) μεταξύ των στοιχείων ενός προγράμματος[5]. Τα κριτήρια όμως μιας τέτοιας επιλογής δεν είναι στόχος του παρόντος. Εφόσον γίνει μια τέτοια επιλογή το επόμενο βήμα είναι η αποτύπωσή της με την μορφή προτύπου. Εκμεταλλευόμενοι την δυνατότητα για τη δημιουργία ιεραρχικών δομών που μας δίνουν τα FASP μπορούμε εύκολα να ορίσουμε σχέση γονέων-παιδιών που να στηρίζονται σε οποιοδήποτε κριτήριο(α) κάθε φορά. Εδώ θα πρέπει να έχουμε στο μυαλό μας πως η ιεραρχία των FASP υπάρχει για να υποδηλώνει την σχέση συνόλου και υποσυνόλου μεταξύ πατέρα και παιδιών αλλά και την, με βάση κάποια κριτήρια, εξειδίκευση διατρέχοντας το δέντρο από τη ρίζα προς τα φύλλα. Βάση ενός τέτοιου προτύπου μπορούμε να δομήσουμε το σύνολο των προτιμήσεων που εξάγουμε κάθε φορά από μια ιστορία χρήσης ενός χρήστη.

Αναλυτικότερα παρέχεται η δυνατότητα να ορίζονται ιεραρχίες με κάθε κόμβο να χαρακτηρίζεται από ένα σύνολο κριτηρίων που θα λειτουργούν σαν στοιχεία ομαδοποίησης των προγραμμάτων και σημασιολογικά θα υποδηλώνουν την δομική σχέση των κόμβων. Για να γίνει αυτό κατανοητό ένα πρότυπο στην περίπτωση των κατηγοριών των προγραμμάτων θα ήταν το εξής:



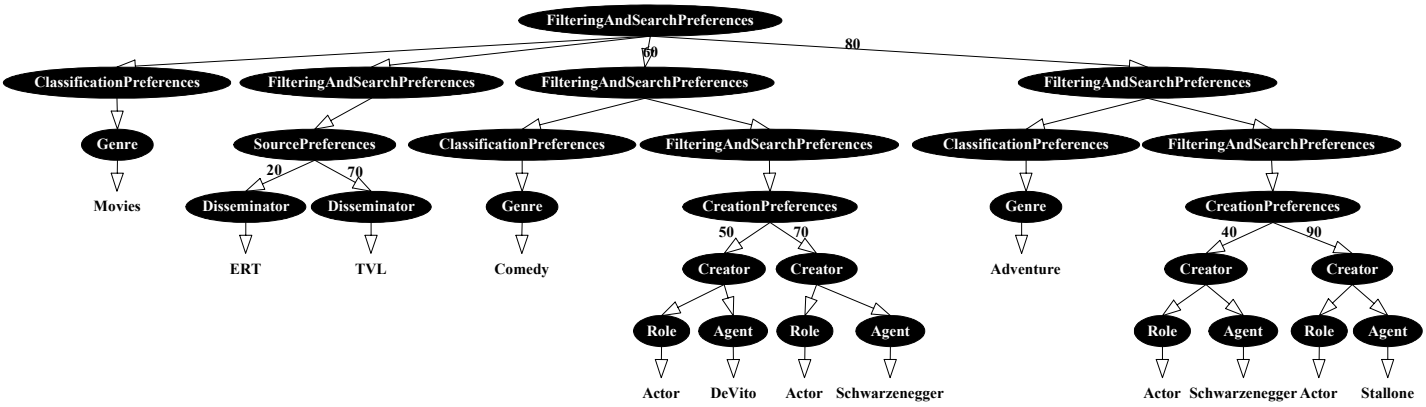
Σχήμα 24: Παράδειγμα Προτύπου FASP με χρήση μόνο των Genres.

Η επιπλέον δυνατότητα που παρέχεται είναι ότι κάτω από κάθε FASP αντί να υπάρχει το genre ή μόνο το genre, μπορεί να υπάρχει οποιοδήποτε ή ακόμα και οποιοδήποτε σύνολο χαρακτηριστικών. Όπως για παράδειγμα ένα σύνολο από ηθοποιούς στην ρίζα και κάποια υποσύνολα αυτών, στα παιδιά.

[5] Όπως με τη χρήση τεχνικών από την περιοχή του Data Mining.

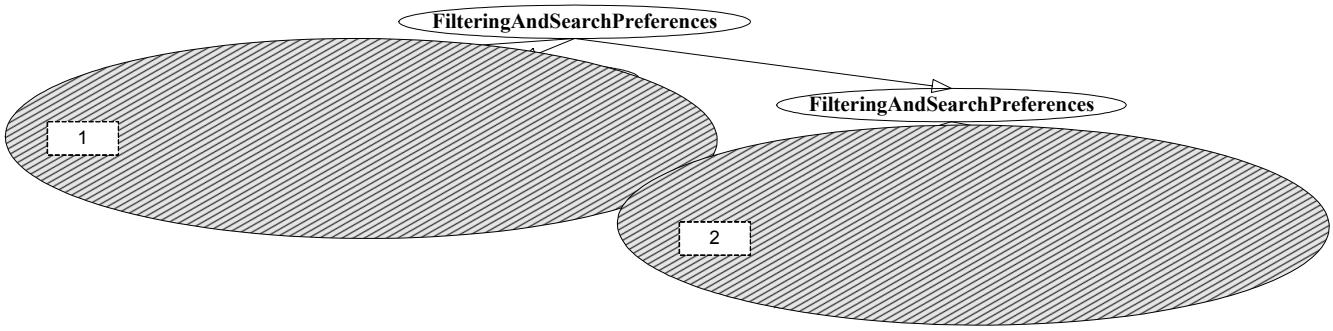
Δημιουργία Ιεραρχικού Προφίλ Βάση Προτύπου

Πρότυπα σαν το παραπάνω μπορούν να υπάρχουν αποθηκευμένα στη βάση και κατ' απαίτηση να χρησιμοποιούνται για την δημιουργία του προφίλ κάποιου χρήστη. Σε μια τέτοια περίπτωση η διαδικασία που ακολουθείται θα είναι: με βάση τα χαρακτηριστικά κάθε κόμβου γίνεται ομαδοποίηση των προγραμμάτων που ικανοποιούν τα χαρακτηριστικά του κόμβου και εύρεση των προτιμήσεων, για όλα τα χαρακτηριστικά των προγραμμάτων, με βάση αυτήν την ομαδοποίηση. Στη συνέχεια δημιουργείται ένα αντίγραφο του προτύπου και οι κόμβοι του λειτουργούν ως οι κύριοι κόμβοι της ιεραρχίας. Στο τελευταίο βήμα κάτω από αυτούς προστίθενται κόμβοι FASP που περιέχουν όλα τα χαρακτηριστικά για τα οποία στο πρώτο βήμα υπολογίστηκε κάποια τιμή προτίμησης. Το προφίλ που θα προκύψει για ένα πρότυπο σαν αυτό του προηγούμενου σχήματος θα μπορούσε για κάποια ενδεικτικά χαρακτηριστικά να είναι το εξής:



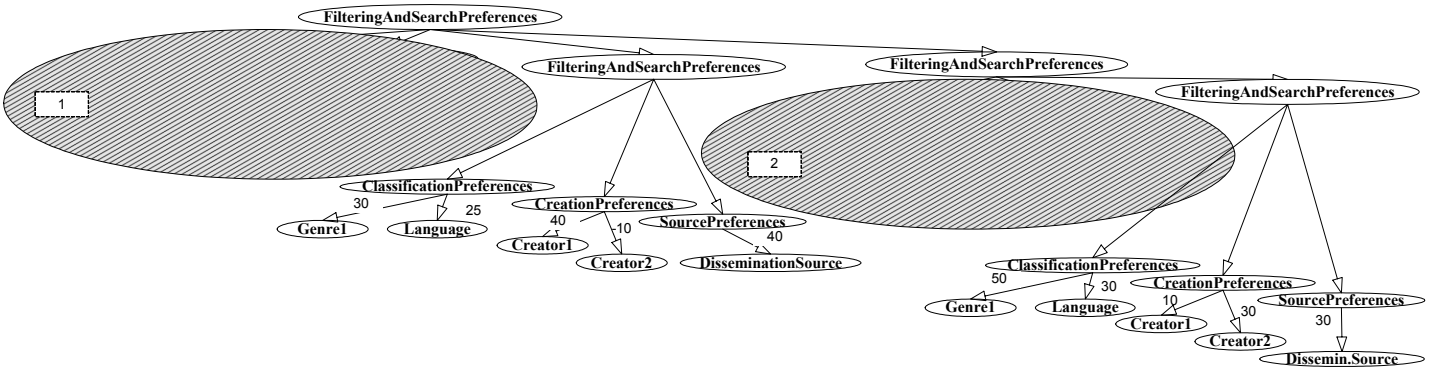
Σχήμα 25: Παράδειγμα Ιεραρχικού προφίλ με βάση το πρότυπο του *σχήματος 24.*

Όπως ήδη έχει διαφανεί, στόχος της χρήσης προτύπων είναι η δυνατότητα για χρησιμοποίηση περισσότερων χαρακτηριστικών ως στοιχεία ομαδοποίησης των προγραμμάτων. Στη συνέχεια θα παρουσιάσουμε ένα τέτοιο, πιο περίπλοκο, πρότυπο, κάποιο ενδεικτικό προφίλ με βάση αυτό το πρότυπο, ενώ κλείνοντας θα παρουσιάσουμε την δυνατότητα για αναπροσαρμογή τέτοιου τύπου προφίλ.
Ας θεωρήσουμε λοιπόν το πρότυπο του επόμενου σχήματος:



Σχήμα 26: Παράδειγμα σύνθετου προτύπου ιεραρχίας από FASP.

Τα χαρακτηριστικά στις περιοχές 1 και 2 είναι αυτά με βάση τα οποία θα γίνει η ομαδοποίηση των προγραμμάτων. Στη συγκεκριμένη περίπτωση η ομαδοποίηση των προγραμμάτων θα γίνει με βάση την κατηγορία τους και τους ηθοποιούς που περιέχουν. Έτσι θα ομαδοποιηθούν τα προγράμματα που ικανοποιούν τα χαρακτηριστικά της περιοχής 1 και κάτω από το FASP πατέρα θα "κρεμαστεί" κόμβος FASP που να περιέχει ότι πληροφορία υπάρχει για τις προτιμήσεις του χρήστη πάνω στα προγράμματα της περιοχής 1. Το ίδιο θα γίνει και για την περιοχή 2 και τελικά το προφίλ που θα προκύψει θα είναι τις εξής μορφής:

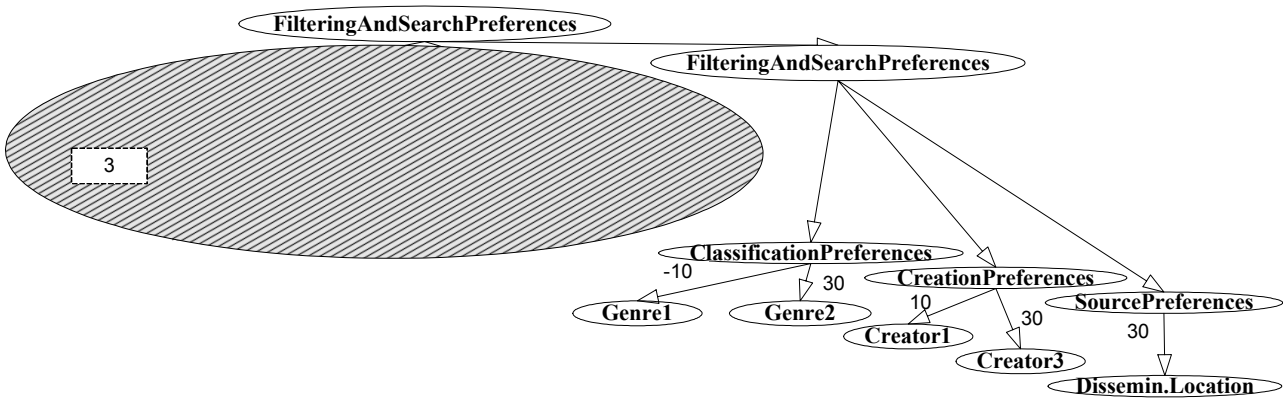


Σχήμα 27: Παράδειγμα Ιεραρχικού Προφίλ με βάση το πρότυπο του *σχήματος 26.*

Οι κόμβοι που έχουν προστεθεί περιέχουν όπως αναφέραμε τις προτιμήσεις του χρήστη για όλα τα προγράμματα που πληρούν τα χαρακτηριστικά των γραμμοσκιασμένων περιοχών. Τα χαρακτηριστικά που έχουν προστεθεί είναι απλά ενδεικτικά για τους λόγους της παρουσίασης.

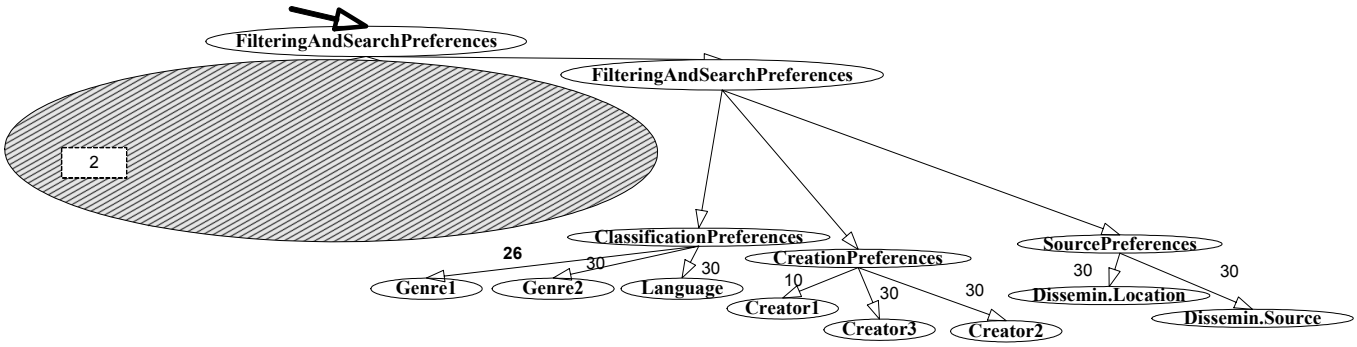
Αναπροσαρμογή Ιεραρχικού Προφίλ Βάση Προτύπου

Όπως και στην ανάλογη περίπτωση του ιεραρχικού προφίλ με βάση την κατηγορία των προγραμμάτων, θα πρέπει πρώτα να έχουν δημιουργηθεί δύο προφίλ με βάση το ίδιο πρότυπο. Αν θεωρήσουμε πως το παλιό είναι αυτό του προηγούμενου παραδείγματος τότε ένα πιθανό καινούριο θα μπορούσε να είναι το εξής:



Σχήμα 28: FASP βάση του προτύπου του *σχήματος 26* που θα χρησιμοποιηθεί για την αναπροσαρμογή του προφίλ στο *σχήμα 27.*

Παρατηρούμε πως η περιοχή 3 του σχήματος .. ταιριάζει με τα κριτήρια της περιοχής 2 του σχήματος ... Αφού γίνει εύρεση αυτής της αντιστοιχίας η ακολουθεί η διαδικασία του συνδυασμού των προτιμήσεων που περιέχουν. Το FASP που θα προκύψει θα είναι αυτό του σχήματος 10 προσαρμοσμένο. Εδώ φαίνεται μόνο ο κόμβος στον οποίο θα υπάρχουν αλλαγές:



Σχήμα 29: Τμήμα του FASP του *σχήματος 26* που αναπροσαρμόστηκε με βάση το FASP *σχήμα 28.*

Παρατηρούμε την προσαρμογή των τομών προτίμησης χαρακτηριστικών που ήταν κοινά και την απλή προσθήκη εκείνων που εμφανίστηκαν για πρώτη φορά.

Υποσύστημα Αξιολόγησης Εκτιμήσεων(AEEM)

Πρόκειται για το υποσύστημα που στόχο έχει την αξιολόγηση των εκτιμήσεων που προκύπτουν από τα υπόλοιπα μέρη του συστήματος. Συγκεκριμένα στόχος του είναι η υποστήριξη μηχανισμών και μετρικών που θα επιτρέπουν την εκτίμηση της απόδοσης των διάφορων μηχανισμών δημιουργίας προφίλ των χρηστών.

Αφετηρία κάθε φορά είναι μια λίστα από προγράμματα μαζί με την εκτίμηση προτίμησης για καθένα από αυτά, σύμφωνα με της προτιμήσεις κάθε χρήστη. Ταυτόχρονα εξασφαλίζουμε για την ίδια ομάδα προγραμμάτων την ρητή-πραγματική προτίμηση του χρήστη. Το παρόν υποσύστημα είναι υπεύθυνο για την αξιολόγηση των προβλέψεων για τις προτιμήσεις του χρήστη συγκρίνοντας τις προβλεπόμενες προτιμήσεις με τις πραγματικές. Στο υπόλοιπο της παραγράφου θα παρουσιάσουμε μια σειρά από μετρικές που έχουν υλοποιηθεί για το παρόν υποσύστημα, ενώ την εφαρμογή τους θα την περιγράψουμε κατά την πειραματική διαδικασία που ακολουθεί στο επόμενο κεφάλαιο.

Σαν πρώτο μέτρο σύγκρισης των προβλέψεων με τις πραγματικές προτιμήσεις των χρηστών υλοποιήθηκε ο μέσος όρος σφάλματος(Mean Average Error, MAE). Αναλυτικότερα πρόκειται για τον υπολογισμό του μέσου όρου των αποκλίσεων των εκτιμημένων τιμών

προτίμησης, για τα προγράμματα, από τις πραγματικές. Ο τύπος υπολογισμού για το MAE είναι ο:

$$MAE = \frac{\sum_{\langle u,i \rangle} \left| P(u,i) - R_{u,i} \right|}{n}$$

όπου *P(u,i)* η πρόβλεψη της τιμής προτίμησης του χρήστη u για το πρόγραμμα i, R η αντίστοιχη ρητή τιμή προτίμησης του χρήστη για το ίδιο πρόγραμμα και *n* το πλήθος των προγραμμάτων για τα οποία έγινε πρόβλεψη.

Ένα δεύτερο μέτρο αξιολόγησης που χρησιμοποιήθηκε ήταν ο υπολογισμός των σφαλμάτων στην εκτίμηση των προγραμμάτων που εμφανίζονται ως τα πιο ενδιαφέροντα για τον χρήστη. Πιο συγκεκριμένα τα προγράμματα κατατάσσονται με βάση την τιμή προτίμησής τους. Στη συνέχεια επιλέγονται αυτά που βρίσκονται στην κορυφή της κατάταξης των προβλέψεων και στην κορυφή της κατάταξης των ρητών προτιμήσεων. Η αξιολόγηση γίνεται με βάση την τομή αυτών των δύο συνόλων. Σύμφωνα με τα παραπάνω όσο πιο πολλά είναι τα κοινά προγράμματα τόσο καλύτερη εκτίμηση των προγραμμάτων που προτιμά ο χρήστης έγινε. Αυτό αποτελεί και σημαντικό κριτήριο για την αξιολόγηση ενός συστήματος που στοχεύει στο να προτείνει ενδιαφέροντα στον χρήστη προγράμματα. Τέλος θα πρέπει να σημειώσουμε πως για την επιλογή των κορυφαίων στην κατάταξη προγραμμάτων τίθεται ένα όριο στο πλήθος τους όπως για παράδειγμα το 20% επί των συνολικών προγραμμάτων που εμφανίζονται στην κατάταξη.

Το τρίτο κριτήριο αξιολόγησης που χρησιμοποιείται είναι τα ποσοστά ανάκτησης(recall) και ακρίβειας(precision) στην πρόβλεψη των προγραμμάτων. Συγκεκριμένα γίνεται πάλι κατάταξη των προγραμμάτων σύμφωνα με την τιμή προτίμησης που τους έχει αποδοθεί το σύνολο αυτών που έχουν τις υψηλότερες βαθμολογίες της κλίμακας συγκρίνεται με το σύνολο αυτών που είχαν αξιολογηθεί με τις υψηλότερες βαθμολογίες από τον χρήστη.

Όλα τα παραπάνω αποτελούν συνήθεις πρακτικές για την αξιολόγηση της απόδοσης συστημάτων. Η εφαρμογή τους στην περίπτωση του δικού μας συστήματος περιγράφεται αναλυτικά στο επόμενο καφάλαιο που αφορά τη πειραματική διαδικασία που ακολουθήθηκε.

Κεφάλαιο 5: *Πειραματική Διαδικασία*

Η παρούσα εργασία ξεκίνησε για να καλύψει μια σειρά από στόχους οι οποίοι και αναλύθηκαν στα πρώτα κεφάλαια. Η υλοποίηση που πραγματοποιήθηκε για να καλύψει της προδιαγραφές του συστήματος ακολούθησε μια σειρά από ανάγκες που προκύπτουν σε ένα σύστημα διαχείρισης προφίλ χρηστών σε περιβάλλον ψηφιακής τηλεόρασης. Και πρόκειται για ένα νέο περιβάλλον, που στην μορφή που περιγράφεται από το πρόγραμμα up-TV, κάνει τα πρώτα του βήματα, τουλάχιστον σαν ένα πιο ευρείας χρήσης σύστημα. Για αυτό το λόγο ακόμα και οι ίδιες του οι ανάγκες βρίσκονται σε στάδιο ορισμού και εξέλιξης. Θα πρέπει λοιπόν να επικεντρώσουμε στην ανάγκη που ήρθε να καλύψει η παρούσα εργασία η οποία δεν είναι άλλη από την όσο πιο εύκολη και γρήγορη πρόσβαση του χρήστη στο διαθέσιμο περιεχόμενο.

Ο ρόλος του τελικού συστήματος, όπως αυτό ορίστηκε και υλοποιήθηκε είναι αυτός του πράκτορα που αυτόματα εντοπίζει το ενδιαφέρον, για τον χρήστη περιεχόμενο και του το προτείνει. Όπως είδαμε μια σειρά από υποσυστήματα υλοποιήθηκαν για να επιτελέσουν τον τελικό στόχο. Σε όλα υπάρχουν παράμετροι οι οποίες και παίζουν καθοριστικό ρόλο για την απόδοση του συστήματος. Και μιλώντας για απόδοση, αναφερόμαστε στην ικανότητα του συστήματος να εντοπίζει τα ενδιαφέροντα του χρήστη, να τα αποτυπώνει στο προφίλ του και στη συνέχεια να τα χρησιμοποιεί ώστε να εντοπίζει ανάμεσα στο διαθέσιμο περιεχόμενο τα προγράμματα εκείνα που πραγματικά ενδιαφέρουν τον χρήστη.

Μια πειραματική διαδικασία θα έπρεπε, λοιπόν, να είναι σε θέση να ελέγξει ακριβώς τα στοιχεία του συστήματος που περιγράφηκαν παραπάνω. Για το σκοπό αυτό στα πλαίσια της εργασίας διεξήχθησαν μια σειρά πειραμάτων με αντικείμενο τον καθορισμό των παραμέτρων και τον έλεγχο της απόδοσης των υποσυστημάτων που υλοποιήθηκαν.

Η εύρεση πλήθους πραγματικών χρηστών και η συλλογή δεδομένων σε πραγματικές συνθήκες είναι διαδικασία χρονοβόρα και απαιτητική σε βαθμό που ξεφεύγει από τις δυνατότητες της παρούσας εργασίας. Παρ' όλα αυτά κρίνοντας πολύτιμη τη διεξαγωγή πειραμάτων για τον έλεγχο του συστήματος και την εξαγωγή συμπερασμάτων αποφασίσαμε την διεξαγωγή πειραμάτων χρησιμοποιώντας υπάρχοντα πειραματικά δεδομένα. Μετά από σχετική έρευνα αποφασίστηκε η χρήση των δεδομένων που είχαν συλλεγεί στο πρόγραμμα GroupLens.

Πρόκειται για ένα σύνολο από 943 χρήστες και τις ρητές προτιμήσεις τους για μια λίστα από ταινίες. Καθένας από αυτούς έχει δώσει βαθμό στις ταινίες σε μια κλίμακα από 1-5 όπου το 1 υποδηλώνει την χαμηλότερη προτίμηση και το 5 την υψηλότερη. Επίσης καθένας έχει βαθμολογήσει τουλάχιστον 20 ταινίες ενώ το σύνολο των βαθμολογιών είναι 100.000. Τέλος, όλες οι ταινίες είναι 1648 και ο τίτλος τους είναι το μόνο διαθέσιμο στοιχείο για αυτές. Αυτά ήταν τα διαθέσιμα δεδομένα. Πρώτος στόχος μας ήταν να φέρουμε τα δεδομένα σε μορφή συμβατή με το σύστημα που έχουμε υλοποιήσει. Για αυτό το σκοπό ήταν απαραίτητη η εύρεση των μεταδομένων των προγραμμάτων και η αποθήκευσή τους στην TVA συμβατή βάση του συστήματος. Η συλλογή των μεταδεδομένων έγινε μέσω της ιστοσελίδας του www.imdb.com που διαθέτει πληροφορία για ένα πολύ μεγάλο σύνολο από ταινίες σε μορφή html. Για τις ταινίες που μας ενδιέφεραν τα μεταδεδομένα που συλλέξαμε ήταν η κατηγορία/ες τους(genre), οι γλώσσα τους, η χώρα προέλευσής τους, οι ηθοποιοί, οι συγγραφείς και οι σκηνοθέτες και μια λίστα από λέξεις κλειδιά για την κάθε μια. Μετά τη συλλογή των δεδομένων ακολούθησε η μετατροπή τους και η αποθήκευσή τους στη βάση του συστήματος. Στην ίδια βάση προστέθηκαν και οι 943 χρήστες για τους οποίους διαθέταμε της βαθμολογίες τους. Ακολούθησε ο καθορισμός και η διεξαγωγή μιας σειράς πειραμάτων τα οποία περιγράφουμε στη συνέχεια μαζί με την παρουσίαση των αποτελεσμάτων τους.

5.1 Ανάλυση Πειραμάτων

Κύριο μέσο διάκρισης των προγραμμάτων είναι τα μεταδεδομένα τους. Με λίγα λόγια όλα εκείνα τα χαρακτηριστικά που αποτελούν το κάθε πρόγραμμα παίζουν και τον ρόλο των κριτηρίων βάση των οποίων επιλέγει ο χρήστης τα προγράμματα που τον ενδιαφέρουν. Στόχος μας λοιπόν θα είναι έχοντας μια ομάδα προγραμμάτων με τις προτιμήσεις ενός χρήστη για αυτά, να προσδιορίσουμε με όσο πιο αποδοτικό τρόπο την προτίμηση του για συγκεκριμένα χαρακτηριστικά που εμφανίζονται στα προγράμματα αυτά. Στη συνέχεια θα ασχοληθούμε με τον υπολογισμό των τιμών προτίμησης για τα χαρακτηριστικά που εμφανίζονται στα προγράμματα που έχει 'καταναλώσει' ο χρήστης. Θα θεωρήσουμε πως για τα πρόγραμμα αυτά υπάρχει είτε ρητή, από τον χρήστη, είτε αυτόματη, από το σύστημα, ένδειξη της σχετικής προτίμησης που τυγχάνουν αυτά από τον χρήστη. Με βάση αυτό σκοπός μας είναι έχοντας μια λίστα από προγράμματα, την προτίμηση του χρήστη για αυτά και τα μεταδεδομένα των προγραμμάτων, να αναπτύξουμε μια μεθοδολογία που θα μας επιτρέπει την εύρεση της σχετικής αξίας κάθε χαρακτηριστικού των προγραμμάτων για τον χρήστη και τον αντικατοπτρισμό της αξίας αυτής με μια τιμή προτίμησης στο προφίλ του χρήστη για το αντίστοιχο χαρακτηριστικό. Στο κείμενο που ακολουθεί θα παρουσιάσουμε μια σειρά από προσεγγίσεις, για την επίτευξη του παραπάνω στόχου, προσπαθώντας να φτάσουμε εξελικτικά σε πιο αποδοτικές μεθόδους.

Θα ξεκινήσουμε από τις πιο απλές προσεγγίσεις και αφού εντοπίσουμε τα τυχόν ελαττώματα τους θα προσπαθήσουμε να βελτιώσουμε το τελικό αποτέλεσμα εισάγοντας πιο σύνθετες έννοιες.

5.1.1 ***Υπολογισμός προτίμησης χαρακτηριστικών των προγραμμάτων:*** Μια πρώτη προσέγγιση

Όταν έχουμε μια λίστα από προγράμματα για κάποιον χρήστη τότε αυτομάτως έχουμε και μια λίστα από χαρακτηριστικά που περιέχουν αυτά τα προγράμματα όπως, ηθοποιοί, λέξεις-κλειδιά, χώρες κ.τ.λ. Σε πολλές περιπτώσεις μερικά χαρακτηριστικά είναι κοινά για κάποια προγράμματα. Ο πιο απλός, ίσως, τρόπος για να καταλήξουμε σε μια εκτίμηση της προτίμησης του χρήστη για κάθε χαρακτηριστικό είναι παίρνοντας τον μέσο όρο των τιμών

προτίμησης των προγραμμάτων που περιέχουν το αντίστοιχό χαρακτηριστικό. Με αυτόν το τρόπο υποθέτουμε πως η τιμή προτίμησης για ένα πρόγραμμα αποτελεί και την ουσιαστική ένδειξη για την προτίμηση όλων των χαρακτηριστικών του. Σύμφωνα με αυτά σε κάθε χαρακτηριστικό η τιμή προτίμησης που θα πρέπει να αποδοθεί θα είναι: $P_i = \frac{\sum_t w_{it}}{n_f}$. Όπου P_i η τιμή προτίμησης για το χαρακτηριστικό *i*, W_{it} η τιμή προτίμησης του χαρακτηριστικού *i* στο πρόγραμμα *t* και n_f το πλήθος των προγραμμάτων που έχει δει ο χρήστης και περιείχαν το χαρακτηριστικό.

5.1.2 *Ανάλυση επίδρασης συχνότητας εμφάνισης χαρακτηριστικών*

Με μια προσεκτικότερη όμως ματιά θα διαπιστώσουμε μια σημαντική διαφορά στην συχνότητα που εμφανίζεται κάθε χαρακτηριστικό σε σχέση πάντα με το πλήθος των προγραμμάτων που έχει δει ο χρήστης. Σύμφωνα με αυτό μια αδιάκριτη αντιμετώπιση των χαρακτηριστικών φαίνεται να «αδικεί» την διαφορετική πληροφορία που υπάρχει για διαφορετικά χαρακτηριστικά. Συγκεκριμένα θα μπορούσε κανείς να υποθέσει πως τα χαρακτηριστικά που εμφανίστηκαν περισσότερες φορές στα προγράμματα του χρήστη έχουν κάποια μεγαλύτερη βαρύτητα για αυτόν. Επιπλέον όταν έχω προτίμηση περισσότερες από μία φορές για ένα χαρακτηριστικό τότε θα μπορούσαμε ίσως να δεχτούμε πως κάτι τέτοιο αποτελεί ισχυρότερη ένδειξη για το τι πραγματικά συμβαίνει για το χαρακτηριστικό σε αντιδιαστολή πάντα με το να είχαμε μόνο μια εμφάνιση του χαρακτηριστικού. Αυτό θα ήταν εύκολο να το αποτυπώσουμε Από μαθηματικής άποψης στον τύπο:

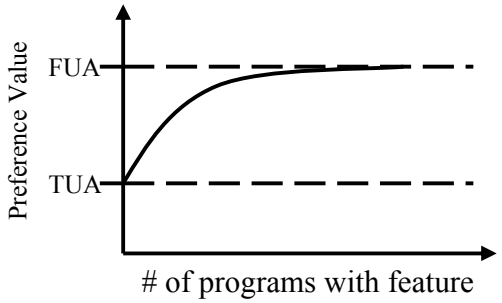
$$P_i = \frac{\sum_t w_{it}}{n_{tot}}$$

Όπου P_i η τιμή προτίμησης για το χαρακτηριστικό *i*, W_{it} η τιμή προτίμησης του χαρακτηριστικού *i* στο πρόγραμμα *t* και *n* το πλήθος όλων των προγραμμάτων που έχει δει ο χρήστης. Με αυτόν το τρόπο όσο περισσότερες φορές εμφανίζεται ένα χαρακτηριστικό τόσο μεγαλύτερη αξία θα αποκτά για τον χρήστη. Τι συμβαίνει όμως όταν ένα χαρακτηριστικό όντως εμφανίστηκε πολλές φορές αλλά όλες τις φορές είχε χαμηλή τιμή

προτίμησης; Σύμφωνα με την παραπάνω προσέγγιση η προτίμηση του θα είναι μεγάλη ανεξάρτητα από την τιμή προτίμησης με την οποία εμφανίστηκε και επιπλέον ακόμα μεγαλύτερη και από ένα χαρακτηριστικό που είχε μεγάλη τιμή προτίμησης αλλά εμφανίστηκε πολύ λιγότερες φορές. Αυτά είναι αρκετά για να οδηγηθούμε σε μια πιο σύνθετη προσέγγιση που θα λαμβάνει υπόψη τη συχνότητα εμφάνισης του χαρακτηριστικού αλλά θα διατηρεί μια 'δίκαια' αντιμετώπιση απέναντι στα χαρακτηριστικά με διαφορετικές τιμές προτίμησης. Σε αυτό το σημείο δεν θα πρέπει να ξεχάσουμε την υπόθεση πως δεν έχω αξιόπιστη ένδειξη για χαρακτηριστικά που εμφανίστηκαν ελάχιστες ή μόνο μία φορά.

5.1.3 *Μελέτη επίδρασης πλήθους βαθμολογημένων προγραμμάτων*

Για κάθε τύπο χαρακτηριστικών και για κάθε χρήστη υπάρχει μια μέση τιμή προτίμησης που προκύπτει από όλα τα προγράμματα που έχει δει ο χρήστης. Ας την ονομάσουμε TUA(Total User Average). Η απλούστερη πρόβλεψη λοιπόν, για το πόσο θα αρέσει ένα άγνωστο πρόγραμμα στο χρήστη είναι TUA. Αυτή η πρόβλεψη δεν λαμβάνει υπόψη της ότι τα πρόγραμμα που έχει δει ο χρήστης και έχουν ένα συγκεκριμένο χαρακτηριστικό, μπορεί να έχουν μια δραστικά διαφορετική μέση τιμή βαθμολογίας από τη συνολική μέση τιμή βαθμολογίας(TUA) των ταινιών που έχει δει ο χρήστης. Επιπλέον υπάρχει και η μέση τιμή προτίμησης για κάθε μεμονωμένο χαρακτηριστικό που προκύπτει μόνο από τα προγράμματα που περιέχουν το συγκεκριμένο χαρακτηριστικό και είναι αυτή που είδαμε παραπάνω. Ας την ονομάσουμε FUA(Feature User Average). Αν λάβουμε υπόψη τα παραπάνω συμπεράσματά μας τότε το FUA θα μπορούσε να θεωρηθεί αξιόπιστο μόνο όταν υπάρχει ένας ικανοποιητικός αριθμός προγραμμάτων που να περιέχουν το χαρακτηριστικό. Στις περιπτώσεις που το χαρακτηριστικό εμφανίστηκε ελάχιστες φορές θα μπορούσαμε να κρίνουμε αναξιόπιστο το αντίστοιχο FUA και να του αποδώσουμε το γενικό TUA. Τέλος θα μπορούσε να υπάρχει μια σταδιακή μετάβαση από την μια τιμή στην άλλη όσο αυξάνει ο αριθμός των προγραμμάτων που περιέχουν ένα χαρακτηριστικό. Για να γίνουν τα παραπάνω πιο σαφή παραθέτουμε το παρακάτω σχήμα:



Σχήμα π1: η τιμή προτίμησης τείνει από το TUA στο FUA όσο αυξάνει ο αριθμός των προγραμμάτων που περιείχαν το χαρακτηριστικό.

Η μετάβαση από τη μια τιμή στην άλλη θα μπορούσε ενδεικτικά να ακολουθεί μια μαθηματική καμπύλη της μορφής :

$$\text{TUA} + (\text{FUA} - \text{TUA}) * (1 - \frac{1}{n})$$

όπου n το πλήθος των προγραμμάτων στα οποία παρατηρήθηκε το αντίστοιχο χαρακτηριστικό. Τώρα είναι σαφές πως αν το χαρακτηριστικό έχει εμφανιστεί μόνο μία φορά τότε θα του αποδοθεί η γενική μέση τιμή(TUA). Όσο το n αυξάνει τόσο η τιμή του προσεγγίζει την πραγματική τιμή προτίμησης του, όπως αυτή έχει προκύψει(FUA). Παρ' όλα αυτά δεν θα μπορούσαμε να περιμένουμε ένα πολύ μεγάλο αριθμό εμφάνισης του χαρακτηριστικού μέχρι να του αποδώσουμε την μέση τιμή προτίμησής του. Αυτό γίνεται κατανοητό αν διατυπώσουμε την υπόθεση πως ένας αριθμός προγραμμάτων όπως για παράδειγμα πέντε ή δέκα είναι αρκετά ικανοποιητικός ώστε να έχουμε αξιόπιστη πληροφορία για την προτίμησή του. Θα μπορούσαμε λοιπόν να τροποποιήσουμε την παραπάνω εξίσωση προκειμένου να υπάρχει μια πιο γρήγορη σύγκλιση στην τιμή του FUA. Ενδεικτικά, για τον τελευταίο όρο της εξίσωσης θα μπορούσαμε να έχουμε εναλλακτικά : (1 - $\frac{1}{n+1}$) ή ακόμα και (1 - $\frac{1}{n^2+2}$). Η τελευταία καμπύλη είναι και αυτή που επιλέξαμε μετά από σειρά πειραμάτων για τον έλεγχο της απόδοσης του συστήματος.

Η ετερογένεια όμως των χαρακτηριστικών μας ωθεί να εξετάσουμε την συμπεριφορά κάθε κατηγορίας χαρακτηριστικού χωριστά, αλλά και κάθε χαρακτηριστικού αναλυτικότερα.

5.1.4 *Μελέτη συμπεριφοράς χαρακτηριστικών στο σύστημα*

Θα μπορούσε λοιπόν να παρατηρήσει κανείς πως υπάρχουν χαρακτηριστικά με πολύ μικρή πληθυκότητα δηλαδή που απαριθμούν ένα μικρό σύνολο τιμών οι οποίες και επαναλαμβάνονται σε όλα τα προγράμματα και άλλα που μπορούν να αποτελούνται από ένα πολύ μεγάλο σύνολο τιμών με λίγες φορές εμφάνισης της κάθε μιας. Για παράδειγμα ένα χαρακτηριστικό που εμφανίζεται λίγες φορές στο σύνολο των προγραμμάτων θα πρέπει να διακρίνεται από κάποιο που εμφανίζεται πολύ συχνά αφού για το πρώτο αρκούν λίγες φορές εμφάνισης στα προγράμματα του χρήστη προκειμένου να θεωρηθεί ισχυρό και αντιστρόφως για το δεύτερο. Διατυπώνουμε λοιπόν την υπόθεση ότι για να πάρει ένα χαρακτηριστικό το μέσο όρο των προτιμήσεων που του αντιστοιχεί(FUA) θα πρέπει η συχνότητα εμφάνισής του στα προγράμματα του χρήστη να είναι ανάλογη της συχνότητας εμφάνισης του χαρακτηριστικού στο σύνολο των προγραμμάτων. Για την ικανοποίηση της παραπάνω υπόθεσης αρκεί να αντικαταστήσουμε το n της παραπάνω εξίσωσης με τον λόγο:

$\frac{n/M}{k/N}$ όπου n το πλήθος των προγραμμάτων του χρήστη που περιέχουν το συγκεκριμένο χαρακτηριστικό, M το πλήθος των προγραμμάτων του χρήστη, k το πλήθος από όλα τα προγράμματα που περιέχουν το χαρακτηριστικό και τέλος N το πλήθος όλων των προγραμμάτων. Εδώ σημειώνουμε απλά πως όταν η συχνότητα εμφάνισης του χαρακτηριστικού για τον χρήστη είναι μικρότερη αυτής για το σύνολο (δηλ. Ο λόγος είναι < 1) τότε αρκεί να τον οδηγήσουμε στην μονάδα προκειμένου να μείνουμε σύμφωνοι με τις παραπάνω υποθέσεις που θεωρούν μια τέτοια περίπτωση ως μη αξιόπιστη για την απόδοση στο χαρακτηριστικό του δικού του μέσου όρου προτίμησης(FUA) και άρα προτιμούν την απόδοση στο χαρακτηριστικό του TUA.

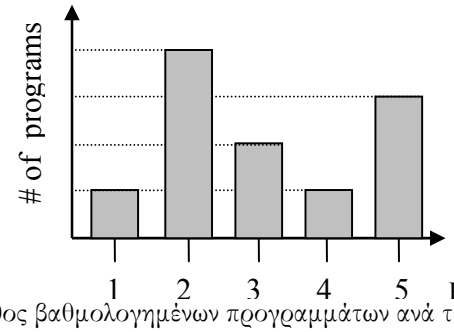
5.1.5 *Μελέτη συμπεριφοράς βαθμολογιών χρήστη με χρήση της εντροπίας*

Αν θελήσουμε να έχουμε μια επιπλέον ένδειξη για την σημαντικότητα του κάθε χαρακτηριστικού για τον χρήστη θα πρέπει να καταφύγουμε στο μέτρο της εντροπίας.

Θυμίζοντας ότι αυτή δίνεται από τον τύπο : $\sum_i^n p_i \log p_i$ ας εξετάσουμε πως θα μπορούσαμε να την εκμεταλλευτούμε στην περίπτωσή μας. Προκειμένου να την ενσωματώσουμε σε όσα ήδη έχουμε πει θα πρέπει πρώτα να την κανονικοποιήσουμε ώστε να κινείται μεταξύ 0 και 1. Θα έχουμε λοιπόν τον εξής τελικό τύπο:

$$\mathrm{E} = \frac{\sum_i^n p_i \log p_i}{\log 1/n}$$

Στην περίπτωσή μας, p_i θα είναι η πιθανότητα κάθε χαρακτηριστικού να πάρει μια από τις τιμές προτίμησης και n το πλήθος των διακριτών τιμών προτίμησης που μπορεί να πάρει ένα χαρακτηριστικό. Για να γίνουν τα παραπάνω πιο σαφή παραθέτουμε το εξής παράδειγμα: Αν 1 ως 5 οι τιμές προτίμησης που μπορεί να πάρει ένα χαρακτηριστικό τότε για κάθε χαρακτηριστικό ενός χρήστη μπορούμε να έχουμε την πληροφορία που φαίνεται στο επόμενο σχήμα:



Σχήμα π2: πλήθος βαθμολογημένων προγραμμάτων ανά τιμή κλίμακας βαθμολογιών

Το πλήθος των παρατηρήσεων για το χαρακτηριστικό αυτό ήταν 11 και τότε p_i = [1/11, 4/11, 2/11, 1/11, 3/11] και E(User, Feature) = Normalized Entropy = $\frac{\sum_i^5 p_i \log p_i}{\log 1/5}$.

Αυτή η πληροφορία μπορεί να αξιοποιηθεί για τον εντοπισμό των σημαντικών τύπων χαρακτηριστικών τόσο για κάθε χρήστη όσο και για όλους του χρήστες.

5.1.6 *Χρήση διασποράς των βαθμολογιών του χρήστη*

Μια τελευταία παρατήρηση που μπορούμε να κάνουμε είναι πως για κάθε χρήστη είναι πολύ πιθανό να παρουσιάζεται μια συγκεκριμένη κατανομή στις αξιολογήσεις που δίνει. Αυτό ίσως θα μπορούσε να αποτελέσει και εναλλακτική επιλογή στη θέση της εντροπίας. Συγκεκριμένα μικρή διασπορά στις βαθμολογίες του χρήστη θα σήμαινε μεγάλη σχέση του χαρακτηριστικού με τη βαθμολογία των προγραμμάτων και αντίστροφα. Επομένως με μεγάλη διασπορά θα θέλαμε πάλι η βαθμολογία του χαρακτηριστικού να προσεγγίζει τον γενικό μέσο όρο και άρα θα μπορούσαμε να έχουμε σαν παράγοντα του αποτελέσματος τον όρο $(1-\frac{Var}{MaxVar})$ όπου Var = $\sum_{i}^{n} |\chi_i - \bar{x}|$ με $\bar{x}$ την μέση προτίμηση του χρήστη στα προγράμματα και $|\chi_i - \bar{x}|$ η απόλυτη τιμή των διαφορών κάθε τιμής προτίμησης από τη μέση ενώ MaxVar η μέγιστη τιμή που μπορεί να πάρει η απόσταση δύο τιμών επί το πλήθος των τιμών, στην περίπτωση αυτή 2*n.

Συνοψίζοντας όλα τα παραπάνω η τελική εξίσωση υπολογισμού της προτίμησης για κάθε χαρακτηριστικό θα είναι :

$$TUA + (FUA - TUA) * \left(1 - \frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}\right) * \left(1-\frac{Var}{MaxVar}\right).$$

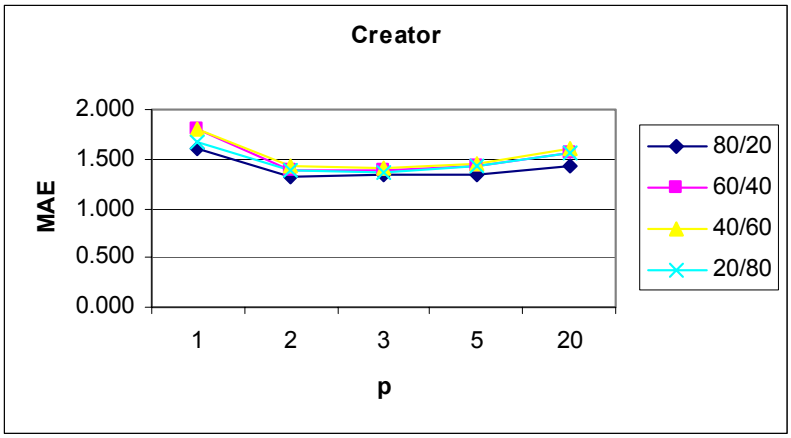
5.2 Παρουσίαση Αποτελεσμάτων

Στη συνέχεια θα περιγράψουμε τα πειράματα που διεξήχθησαν, βασισμένα στις παραπάνω υποθέσεις, μαζί με τα αποτελέσματα που προέκυψαν. Έχοντας ένα σύνολο από 100000 βαθμολογίες χρηστών η διαδικασία που ακολουθήθηκε ήταν η εξής: Σε κάθε πείραμα γίνεται χωρισμός των βαθμολογιών σε δύο σύνολα για κάθε χρήστη. Το πρώτο σύνολο περιέχει βαθμολογίες που χρησιμοποιούνται για τον εντοπισμό των ενδιαφερόντων του χρήστη προκειμένου να κατασκευαστεί το προφίλ και γι' αυτό ονομάζεται σύνολο βαθμολογιών εκπαίδευσης(train set). Το δεύτερο σύνολο περιέχει βαθμολογίες που θα χρησιμοποιηθούν για την αξιολόγηση των προτεινόμενων τεχνικών και γι' αυτό ονομάζεται σύνολο βαθμολογιών αξιολόγησης(test set). Οι βαθμολογίες εκπαίδευσης και οι βαθμολογίες αξιολόγησης για έναν συγκεκριμένο χρήστη αντιστοιχούν σε δύο σύνολα προγραμμάτων που είναι ξένα μεταξύ τους. Εδώ δεν θα πρέπει να παραλείψουμε να αναφέρουμε πως ο συσχετισμός των προφίλ των χρηστών με τα προγράμματα που αντιστοιχούν στις βαθμολογίες αξιολόγησης, γίνεται όπως ήδη έχουμε αναφέρει από το υποσύστημα R-FA με χρήση του p-norm μοντέλου. Για αυτό το λόγο στα πειράματα που ακολουθούν μια από τις παραμέτρους που εξετάζουμε είναι το p για το p-norm μοντέλο και αυτό φαίνεται στα διαγράμματα που παρουσιάζονται κάθε φορά. Στη συνέχεια και καθώς γνωρίζουμε τις ρητές βαθμολογίες αξιολόγησης που έχουν δώσει οι χρήστες στα προγράμματα, χρησιμοποιούμε μια σειρά από μετρικές για αξιολόγηση των μεθοδολογιών που ακολουθήθηκαν κάθε φορά.

5.2.1 Πείραμα: χωρισμός δεδομένων

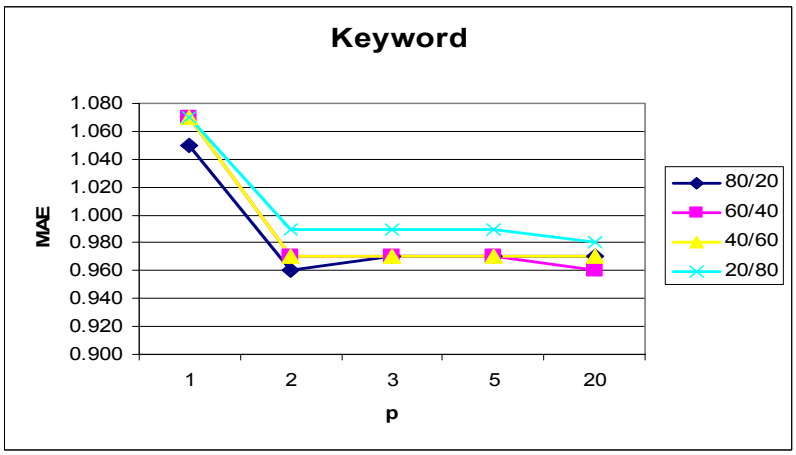
Το πρώτο ερώτημα που προκύπτει είναι το ποιο θα έπρεπε να είναι κάθε φορά το ποσοστό των προγραμμάτων που θα χρησιμοποιηθούν για την κατασκευή των προφίλ και ποιο για την αξιολόγηση. Για να ελέγξουμε τη συμπεριφορά του συστήματος δοκιμάσαμε τέσσερις διαφορετικούς χωρισμούς των δεδομένων με αντίστοιχα ποσοστά 80%-20%, 60%-40%, 40%-60%, 20%-80%. Η δοκιμή έγινε για τον πιο απλό τρόπο εκτίμησης των προτιμήσεων για τα χαρακτηριστικά των προγραμμάτων δηλαδή με τον υπολογισμό του μέσου όρου των προτιμήσεων των προγραμμάτων που περιέχουν το κάθε χαρακτηριστικό. Τα αποτελέσματα που πήραμε φαίνονται στα παρακάτω διαγράμματα. Τα διαγράμματα είναι για το συνδυασμό των χαρακτηριστικών των προγραμμάτων αλλά και για μεμονωμένα

χαρακτηριστικά όπως η κατηγορία, ο ηθοποιός, και οι λέξεις κλειδιά. Τέλος το μέτρο που χρησιμοποιήθηκε ήταν ο μέσος όρος σφάλματος (Mean Average Error-MAE), όπως αυτός ορίστηκε στο κεφάλαιο 4 και στο Υποσύστημα Αξιολόγησης Εκτιμήσεων.



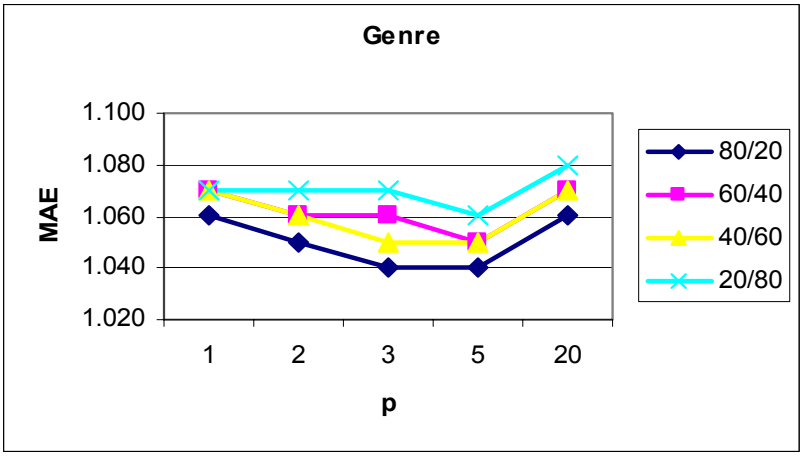
Διάγραμμα 1: Παρουσίαση MAE για προφίλ με χρήση μόνο του Creator. Ως προτίμηση για κάθε χαρακτηριστικό αποδίδεται ο μέσος όρος των προτιμήσεων για τα προγράμματα που το περιέχουν. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, 4 χωρισμοί των δεδομένων σε train set/ test set.

Παρατηρήσεις: η συμπεριφορά του συστήματος είναι παρόμοια για όλες τις περιπτώσεις χωρισμού των δεδομένων. Σχετικά μικρότερο σφάλμα έχουμε για την περίπτωση του 80%-20%(train-test).



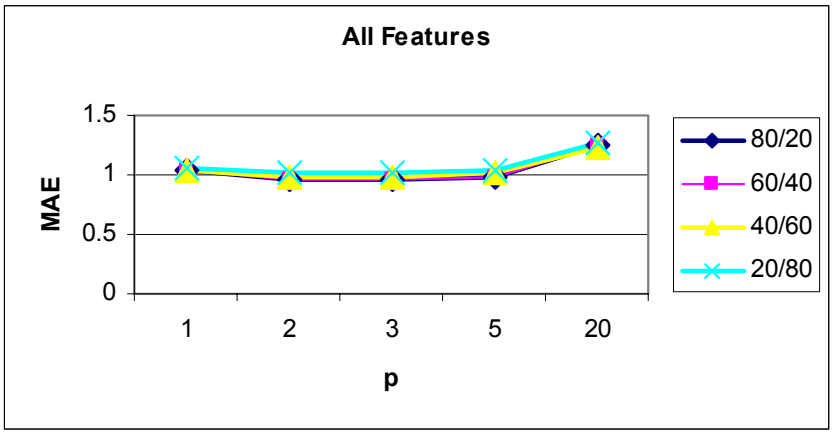
Διάγραμμα 2: Παρουσίαση MAE για προφίλ με χρήση μόνο του Keyword. Ως προτίμηση για κάθε χαρακτηριστικό αποδίδεται ο μέσος όρος των προτιμήσεων για τα προγράμματα που το περιέχουν. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, 4 χωρισμοί των δεδομένων σε train set/ test set.

Παρατηρήσεις: Σε αυτή την περίπτωση παρατηρούμε μεγαλύτερες διαφορές για κάθε περίπτωση χωρισμού των δεδομένων. Εκτός από την περίπτωση του 20%-80% όλες υπόλοιπες σημειώνουν τις μικρότερες τιμές για το MAE. Επίσης παρατηρούμε μια σημαντική βελτίωση για p μεγαλύτερα του 1. Αυτό θα μπορούσε να δικαιολογηθεί από το γεγονός ότι για p=1 το μοντέλο παρουσιάζει πιο αυστηρή συζευκτική συμπεριφορά γεγονός που δεν ευνοεί ένα χαρακτηριστικό σαν το keyword που συμπεριφέρεται καλύτερα για διαζευκτική συμπεριφορά.



Διάγραμμα 3: Παρουσίαση MAE για προφίλ με χρήση μόνο του Genre. Ως προτίμηση για κάθε χαρακτηριστικό αποδίδεται ο μέσος όρος των προτιμήσεων για τα προγράμματα που το περιέχουν. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, 4 χωρισμοί των δεδομένων σε train set/ test set.

Παρατηρήσεις: Στη περίπτωση του Genre παρατηρούμε πως για όλες τις τιμές του p την καλύτερη συμπεριφορά έχουμε για χωρισμό των δεδομένων σε 80% train data, 20% test data.



Διάγραμμα 4: Παρουσίαση MAE για προφίλ με χρήση συνδυασμού των χαρακτηριστικών. Ως προτίμηση για κάθε χαρακτηριστικό αποδίδεται ο μέσος όρος των προτιμήσεων για τα προγράμματα που το περιέχουν. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, 4 χωρισμοί των δεδομένων σε train set/ test set.

Παρατηρήσεις: για το συνδυασμό των χαρακτηριστικών έχουμε ίδια συμπεριφορά για όλους τους εναλλακτικούς χωρισμούς δεδομένων.

Συμπεράσματα

Όπως παρατηρούμε η συμπεριφορά του συστήματος παρουσιάζεται σταθερή με μια ελαφρώς πιο καλή συμπεριφορά για την περίπτωση του 80-20%. Αργότερα που θα χρησιμοποιήσουμε πιο σύνθετες μεθόδους υπολογισμού των προτιμήσεων θα δοκιμάσουμε για άλλη μια φορά τη συμπεριφορά του συστήματος για τα διαφορετικά ποσοστά χωρισμού των δεδομένων. Στο έξης όπου δεν αναφέρεται ο χωρισμός των δεδομένων για το αντίστοιχο πείραμα θα είναι 80% για την κατασκευή του προφίλ και 20% για την αξιολόγηση των εκτιμήσεων.

5.2.2 Συγκριτική παρουσίαση μεθόδων υπολογισμού προτίμησης

Στη συνέχεια και σύμφωνα με όσα αναλύσαμε στην προηγούμενη ενότητα υπάρχουν μια σειρά από διαφορετικούς τρόπους για τον υπολογισμό των προτιμήσεων κάθε χρήστη για τα χαρακτηριστικά των προγραμμάτων. Παρακάτω τους αναφέρουμε συγκεντρωτικά:

1) TUA: ο συνολικός μέσος όρος προτιμήσεων για τα προγράμματα του χρήστη.
2) FUA: ο μέσος όρος των προγραμμάτων που περιέχουν το αντίστοιχο χαρακτηριστικό.
3) TUA + (FUA - TUA) * (1 - $\frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}$): η τιμή βαίνει από το TUA στο FUA όσο

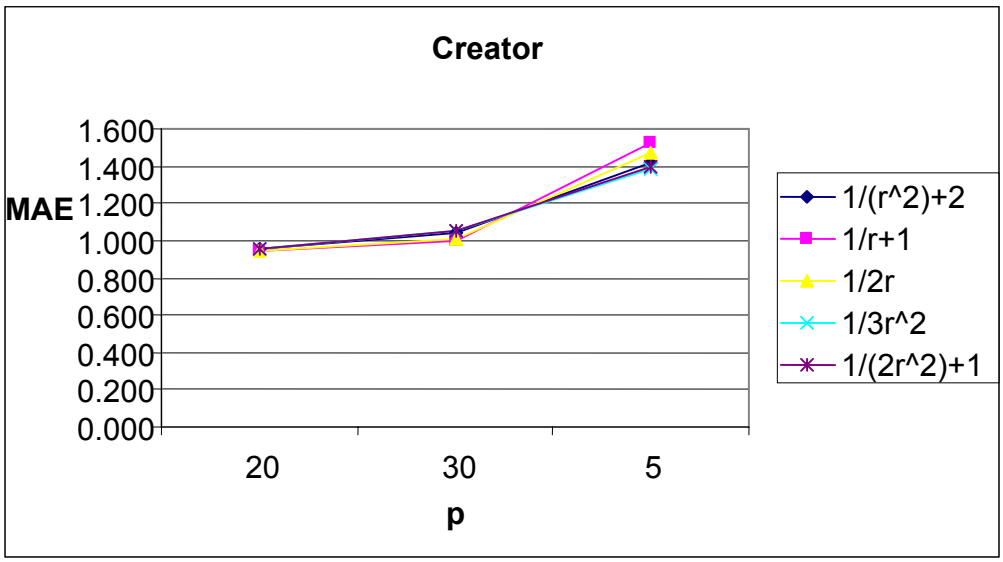
   πιο μεγάλη είναι η συχνότητα εμφάνισης ενός χαρακτηριστικού στα προγράμματα του χρήστη σε σχέση με τη συχνότητα εμφάνισης του ίδιου χαρακτηριστικού στο σύνολο των προγραμμάτων του συστήματος.
4) TUA + (FUA - TUA) * (1 - $\frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}$) * (1 - $\frac{Var}{MaxVar}$): ο τελευταίος όρος

   οδηγεί την τιμή από το TUA στο FUA όσο μικρότερη είναι η διασπορά των βαθμολογιών του χρήστη.
5) TUA + (FUA - TUA) * (1 - $\frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}$) * (1 - $\frac{\sum_{i}^{n} p_i \log p_i}{\log 1/n}$): ο τελευταίος όρος

   οδηγεί την τιμή από το TUA στο FUA όσο μικρότερη είναι η εντροπία των βαθμολογιών του χρήστη.

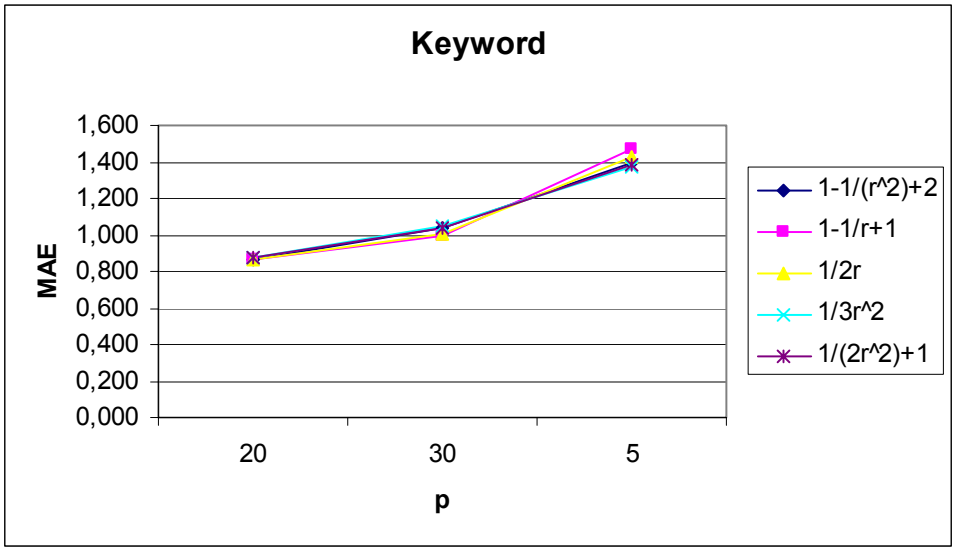
5.2.2.1 **Πείραμα:** Μελέτη καμπύλης σύγκλισης

Πριν παρουσιάσουμε τα συγκριτικά αποτελέσματα των παραπάνω μεθοδολογιών παρουσιάζουμε τη συμπεριφορά του συστήματος για τις διαφορετικές καμπύλες σύγκλισης του όρου $(1 - \frac{1}{x})$.



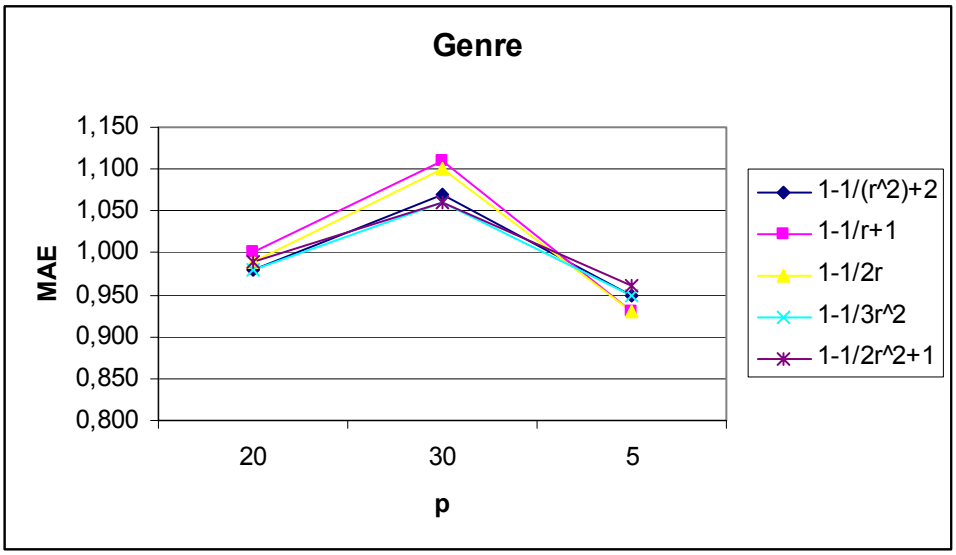
Διάγραμμα 5: Μελέτη καμπύλης σύγκλισης για το Creator. Ο τύπος που χρησιμοποιείται είναι ο TUA+(FUA-TUA)*(1-f), όπου f η εκάστοτε καμπύλη. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, διαφορετικές εξισώσεις σύγκλισης.

Παρατηρήσεις: το MAE είναι παρόμοιο για όλες τι περιπτώσεις. Οι 1/r+1 και 1/2r σημειώνουν κάποιες μεγάλες τιμές σφάλματος για μικρό p.



Διάγραμμα 6: Μελέτη καμπύλης σύγκλισης για το Keyword. Ο τύπος που χρησιμοποιείται είναι ο TUA+(FUA-TUA)*f, όπου f η εκάστοτε καμπύλη. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, διαφορετικές εξισώσεις σύγκλισης.

Παρατηρήσεις: οι καμπύλες παρουσιάζουν την ίδια συμπεριφορά με την περίπτωση του ηθοποιού.



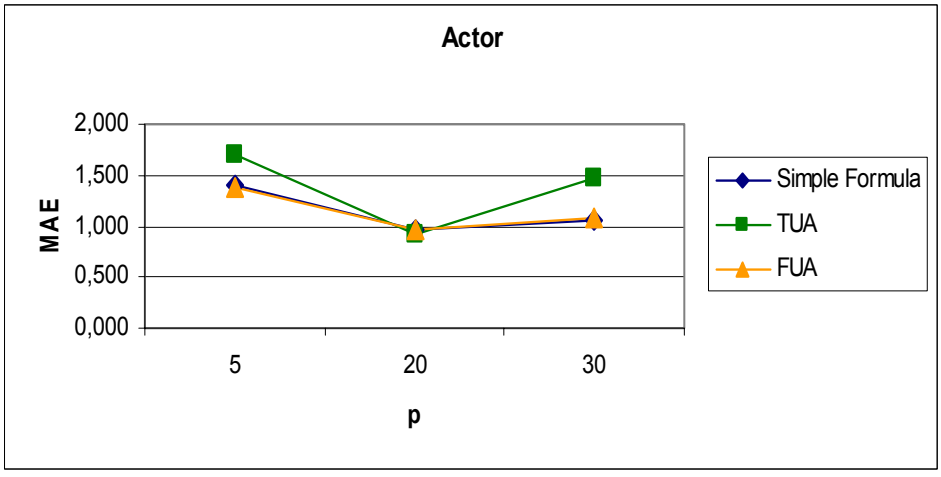
Διάγραμμα 7: Μελέτη καμπύλης σύγκλισης για το Genre. Ο τύπος που χρησιμοποιείται είναι ο TUA+(FUA-TUA)*f, όπου f η εκάστοτε καμπύλη. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, διαφορετικές εξισώσεις σύγκλισης.

Παρατηρήσεις: οι 1/r+1 και 1/2r παρουσιάζουν το μικρότερο σφάλμα, όμως για τα υπόλοιπα p έχουν τη χειρότερη συμπεριφορά.

Συμπεράσματα: Παρατηρούμε πως συμπεριφορά των συναρτήσεων είναι παρόμοια. Τελικά, όπως ήδη έχουμε πει, επιλέξαμε την $(1 - \frac{1}{[n]^2 + 2})$ αφού παρουσιάζει καλύτερη συμπεριφορά για περισσότερες περιπτώσεις.

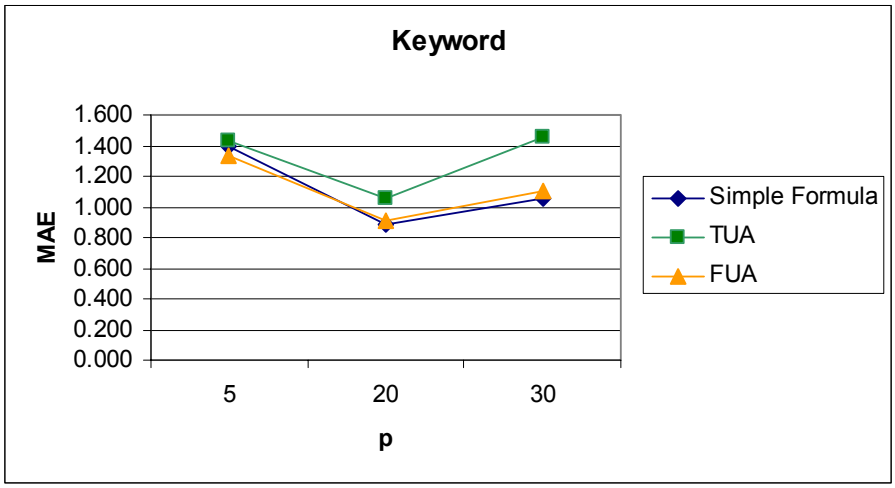
5.2.2.2 Σύγκριση απλού υπολογισμού και υπολογισμού με σύνθετη μέθοδο.

Σύμφωνα με τα παραπάνω ο βασικός όρος της εξίσωσης υπολογισμού των προτιμήσεων για τα διάφορα χαρακτηριστικά θα είναι αυτός της περίπτωσης 3 δηλαδή ο : TUA + (FUA - TUA) * $(1 - \frac{1}{[\frac{n/M}{k/N}]^2 + 2})$. Στα αμέσως επόμενα διαγράμματα συγκρίνουμε τη χρήση αυτού του τρόπου υπολογισμού των τιμών προτίμησης με χρήση απλά είτε του TUA είτε του FUA ως την τιμή προτίμησης. Η σύγκριση γίνεται για τα τρία βασικά χαρακτηριστικά των προγραμμάτων δηλαδή της κατηγορίας(Genre), του ηθοποιού(Actor), και τον λέξεων κλειδιών(Keyword).



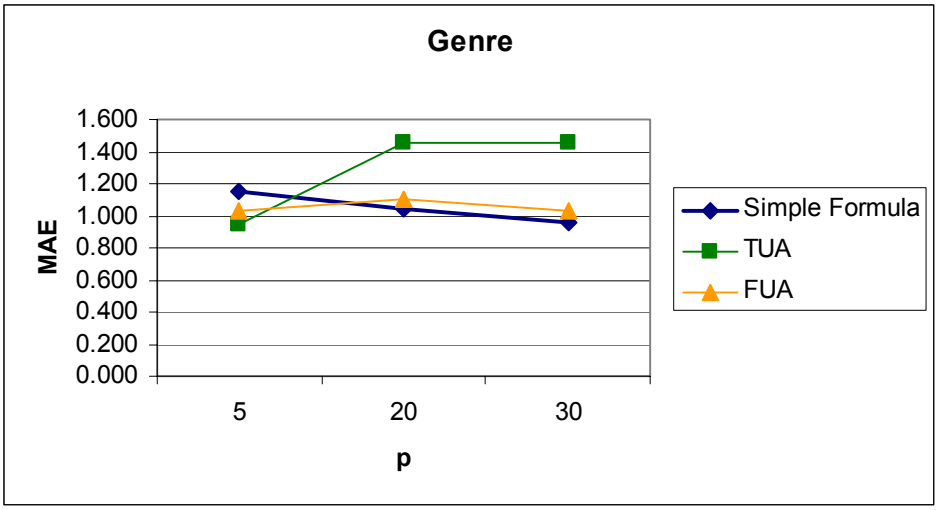
Διάγραμμα 8: Σύγκριση χρήσης καμπύλης σύγκλισης με απλό υπολογισμό προτίμησης για τον Actor. Σαν προτίμηση αποδίδεται το TUA, το FUA και μετά η εξίσωση(**TUA + (FUA - TUA) * (1 -** $\frac{1}{[\frac{n/M}{k/N}]^2 + 2}$**))** με τη καμπύλη σύγκλισης μεταξύ τους. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, τρόπος υπολογισμού τιμής προτίμησης.

Παρατηρήσεις: και οι τρεις περιπτώσεις σημειώνουν χαμηλό MAE για p=20. Για τα υπόλοιπα όμως p το TUA έχει πολύ μεγαλύτερο MAE.



Διάγραμμα 9: Σύγκριση χρήσης καμπύλης σύγκλισης με απλό υπολογισμό προτίμησης για το Keyword. Σαν προτίμηση αποδίδεται το TUA, το FUA και μετά η εξίσωση(**TUA + (FUA - TUA) * (1 -** $\frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}$**)**) με τη καμπύλη σύγκλισης μεταξύ τους. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, τρόπος υπολογισμού τιμής προτίμησης.

Παρατηρήσεις: το TUA παρουσιάζει για όλες τις περιπτώσεις την χειρότερη συμπεριφορά ενώ η χρήση της απλής εξίσωσης παρουσιάζει σχεδόν ίδια συμπεριφορά με το FUA.



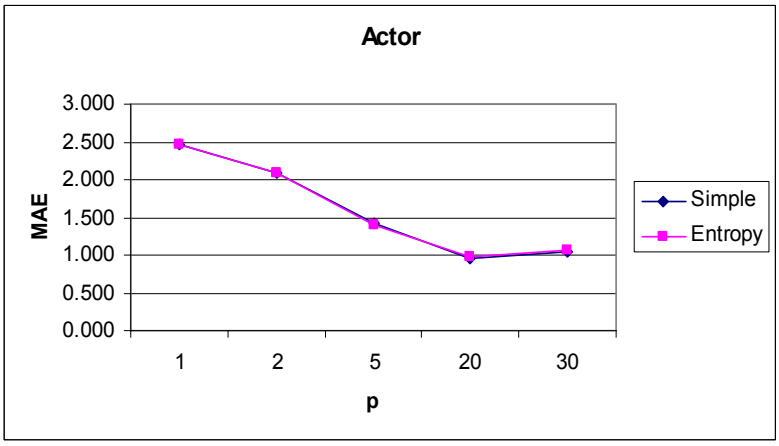
Διάγραμμα 10: Σύγκριση χρήσης καμπύλης σύγκλισης με απλό υπολογισμό προτίμησης για το Genre. Σαν προτίμηση αποδίδεται το TUA, το FUA και μετά η εξίσωση (**TUA + (FUA - TUA) * (1 -** $\frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}$**))** με τη καμπύλη σύγκλισης μεταξύ τους. **Παράμετροι:** διαφορετικά p κατά τη συσχέτιση, τρόπος υπολογισμού τιμής προτίμησης.

Παρατηρήσεις: η συμπεριφορά για το Genre σε σχέση με την επιλογή για την εκτίμηση των προτιμήσεων είναι ίδια με το keyword όμως σε αυτή τη περίπτωση αξίζει να σημειωθεί πως το μικρότερο MAE το έχουμε για p=5 αντί για p=20 που ήταν η καλύτερη περίπτωση για τα άλλα χαρακτηριστικά.

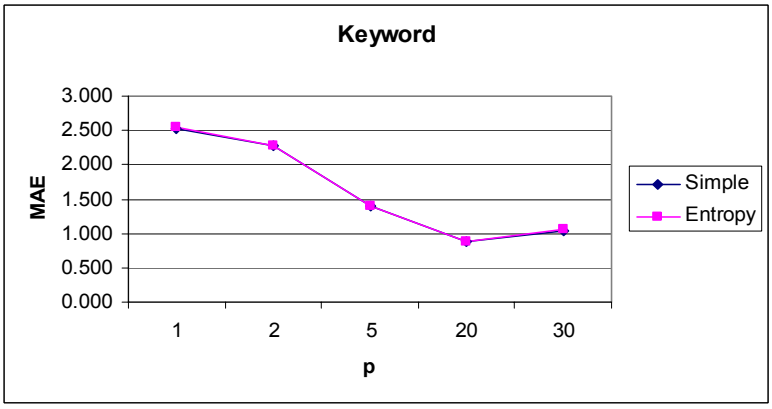
Συμπεράσματα: Αυτό που παρατηρούμε είναι πως η χρήση της συνάρτησης σύγκλισης βελτιώνει σημαντικά τα αποτελέσματα σε σχέση με την απόδοση απλά του γενικού μέσου όρου(TUA) σαν τιμή προτίμησης ενώ υπάρχει μικρή διαφορά από την χρήση της μέσης προτίμησης κάθε χαρακτηριστικού(FUA). Αυτό ήταν και αναμενόμενο αφού η χρήση της συνάρτησης σύγκλισης διαφοροποιεί μόνο τις περιπτώσεις των χαρακτηριστικών που εμφανίστηκαν σε πολύ μικρό αριθμό προγραμμάτων του χρήστη σε σχέση πάντα με τη συχνότητα εμφάνισής τους στο σύνολο των προγραμμάτων.

5.2.2.3 **Πείραμα:** Χρήσης εντροπίας

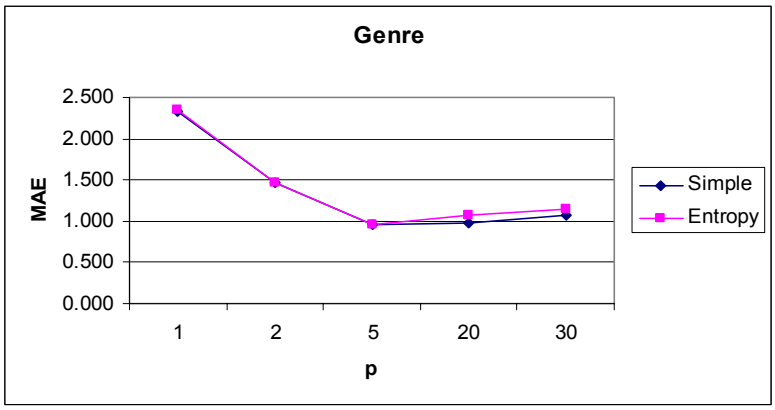
Στα επόμενα διαγράμματα έχουμε προσθέσει και τον τελευταίο όρο της εξίσωσης ($1 - \frac{\sum_{i}^{n} p_i \log p_i}{\log 1/n}$), δηλαδή την εντροπία και κάνουμε σύγκριση της νέας εξίσωσης με τη πιο απλή που χρησιμοποιήσαμε προηγουμένως **(**$\mathrm{TUA} + (\mathrm{FUA} - \mathrm{TUA}) * (1 - \frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2})$**)**:



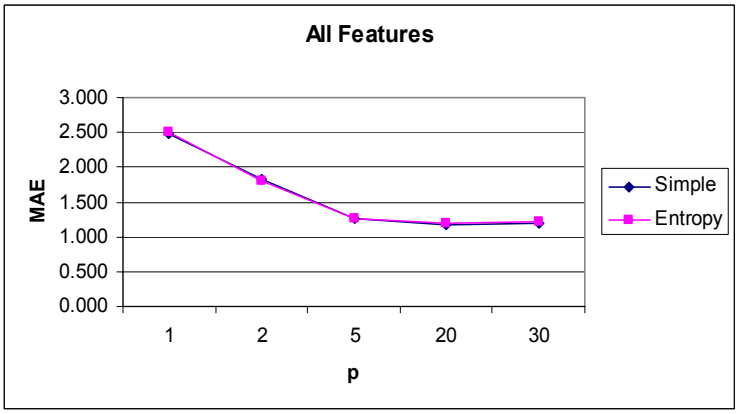
Διάγραμμα 11: Συγκριτικό διάγραμμα καμπύλης με και χωρίς την χρήση της εντροπίας για τον ηθοποιό. Η εντροπία χρησιμοποιείται σαν επιπλέον όρος στην απλή εξίσωση των προηγούμενων πειραμάτων. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με και χωρίς την εντροπία.



Διάγραμμα 12: Συγκριτικό διάγραμμα καμπύλης με και χωρίς την χρήση της εντροπίας για το keyword. Η εντροπία χρησιμοποιείται σαν επιπλέον όρος στην απλή εξίσωση των προηγούμενων πειραμάτων. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με και χωρίς την εντροπία.



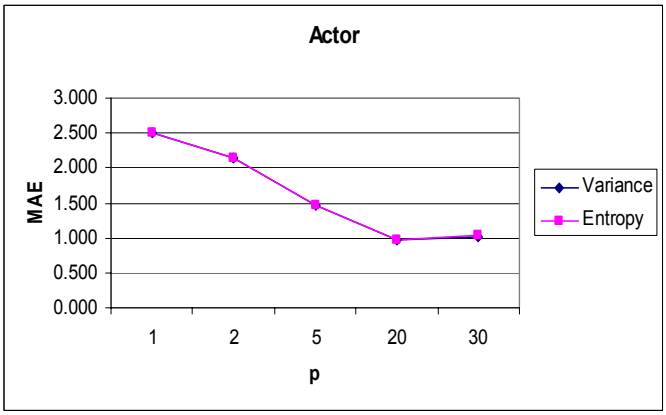
Διάγραμμα 13: Συγκριτικό διάγραμμα καμπύλης με και χωρίς την χρήση της εντροπίας για το Genre. Η εντροπία χρησιμοποιείται σαν επιπλέον όρος στην απλή εξίσωση των προηγούμενων πειραμάτων. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με και χωρίς την εντροπία.



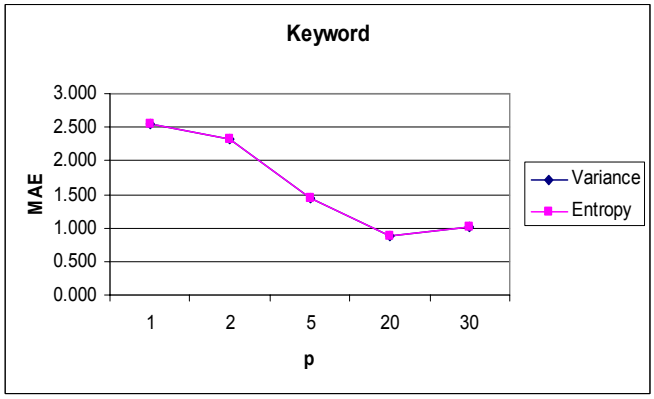
Διάγραμμα 14: Συγκριτικό διάγραμμα καμπύλης με και χωρίς την χρήση της εντροπίας για τον συνδυασμό των χαρακτηριστικών. Η εντροπία χρησιμοποιείται σαν επιπλέον όρος στην απλή εξίσωση των προηγούμενων πειραμάτων. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με και χωρίς την εντροπία.

Συμπεράσματα: Παρατηρούμε πως δεν υπάρχει κάποια σημαντική διαφορά, τουλάχιστον όσον αφορά στο MAE μεταξύ της χρήσης των δύο εξισώσεων. Αυτό θα μπορούσε να σημαίνει πως η συμπεριφορά των βαθμολογιών του χρήστη δεν αποτελεί παράγοντα που να μπορεί να βοηθήσει σε όλες τις περιπτώσεις την εκτίμηση της προτίμησης για τα διάφορα χαρακτηριστικά.

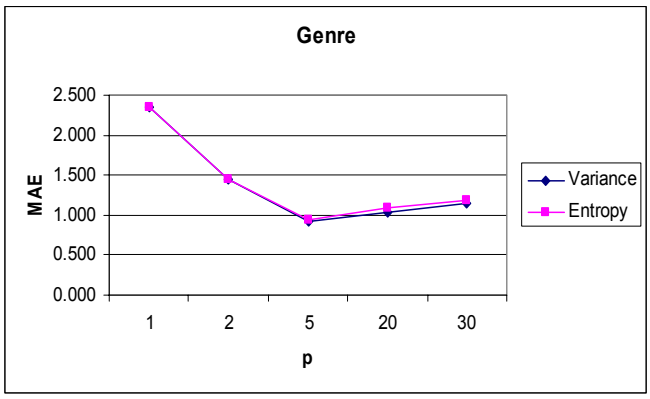
Αμέσως μετά παρουσιάζουμε τα αντίστοιχα διαγράμματα για την εξίσωση με την εντροπία((1 - $\frac{\sum_{i}^{n} p_i \log p_i}{\log 1/n}$)) και την ίδια εξίσωση με το μέτρο της διασποράς των βαθμολογιών των χρηστών (1 - $\frac{Var}{MaxVar}$) στη θέση της εντροπίας.



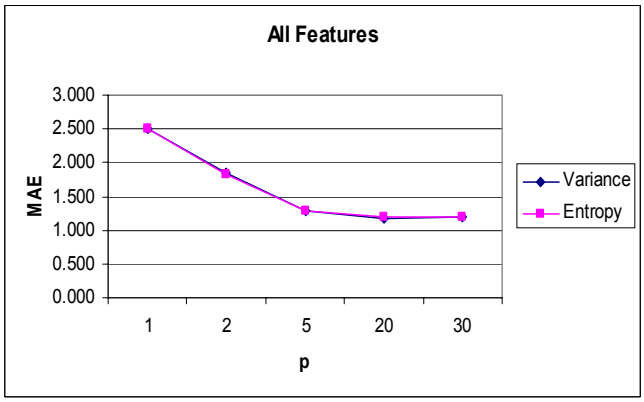
Διάγραμμα 15: Συγκριτικό διάγραμμα με εναλλακτική χρήση εντροπίας και διασποράς βαθμολογιών για τον Actor. Χρησιμοποιούνται εναλλακτικά οι δύο όροι, για την εντροπία και τη διασπορά, στην εξίσωση. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με εντροπία και με διασπορά βαθμολογιών.



Διάγραμμα 16: Συγκριτικό διάγραμμα με εναλλακτική χρήση εντροπίας και διασποράς βαθμολογιών για το Keyword. Χρησιμοποιούνται εναλλακτικά οι δύο όροι, για την εντροπία και τη διασπορά, στην εξίσωση. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με εντροπία και με διασπορά βαθμολογιών.



Διάγραμμα 17: Συγκριτικό διάγραμμα με εναλλακτική χρήση εντροπίας και διασποράς βαθμολογιών για το Genre. Χρησιμοποιούνται εναλλακτικά οι δύο όροι, για την εντροπία και τη διασπορά, στην εξίσωση. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με εντροπία και με διασπορά βαθμολογιών.



Διάγραμμα 18: Συγκριτικό διάγραμμα με εναλλακτική χρήση εντροπίας και διασποράς βαθμολογιών για όλα τα χαρακτηριστικά. Χρησιμοποιούνται εναλλακτικά οι δύο όροι, για την εντροπία και τη διασπορά, στην εξίσωση. **Παράμετροι:** διαφορετικές τιμές του p για την συσχέτιση, εξίσωση με εντροπία και με διασπορά βαθμολογιών.

Συμπεράσματα

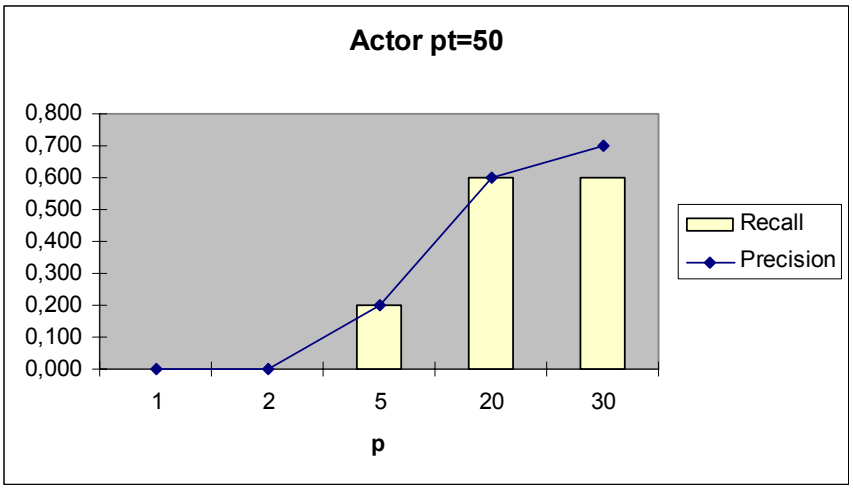
Παρατηρούμε ακριβώς ίδια συμπεριφορά του συστήματος στις δύο περιπτώσεις. Αυτό θα μπορούσε να θεωρηθεί αναμενόμενο αφού και η διασπορά των βαθμολογιών των χρηστών αλλά και η εντροπία αντικατοπτρίζουν τη συμπεριφορά του χρήστη σαν βαθμολογητή. Η διαφορά τους βρίσκεται στη χρήση διαφορετικής μαθηματικής έκφρασης.

Σε όλα τα παραπάνω διαγράμματα χρησιμοποιούμε σαν μέτρο το MAE και παρατηρούμε τη συμπεριφορά κάθε χαρακτηριστικού σε σχέση με τις διαφορετικές τιμές τις παραμέτρου

p. Εδώ θα πρέπει να θυμίσουμε πως το p-norm μοντέλο για μεγάλα p συμπεριφέρεται σαν OR μεταξύ των αντικειμένων που συσχετίζει ενώ για μικρές τιμές του p συμπεριφέρεται σαν πιο αυστηρό Boolean AND. Κάτι τέτοιο θα σήμαινε πως για χαρακτηριστικά σαν τους ηθοποιούς θα περιμέναμε να δούμε καλύτερη απόδοση για διαζευκτική συμπεριφορά ενώ για χαρακτηριστικά σαν τις κατηγορίες των προγραμμάτων θα περιμέναμε καλύτερη απόδοση για συζευκτική συμπεριφορά. Με απλά λόγια ένα AND μεταξύ ενός συνόλου ηθοποιών δεν θα περιμέναμε να συσχετιστεί εύκολα με κάποια ταινία. Από την άλλη πλευρά ένα AND μεταξύ κάποιων κατηγοριών προγραμμάτων αποτελεί σημαντικό κριτήριο συσχετισμού μεταξύ ταινιών που ανήκουν σε ενδιαφέρουσες για τον χρήστη κατηγορίες. Στα αποτελέσματα που παρουσιάσαμε παρατηρούμε πως το Genre παρουσιάζει την καλύτερη συμπεριφορά για p=5 ενώ για τα υπόλοιπα χαρακτηριστικά η καλύτερη συμπεριφορά παρουσιάζεται για p=20 γεγονός που έρχεται σε πλήρη συμφωνία με τα παραπάνω.

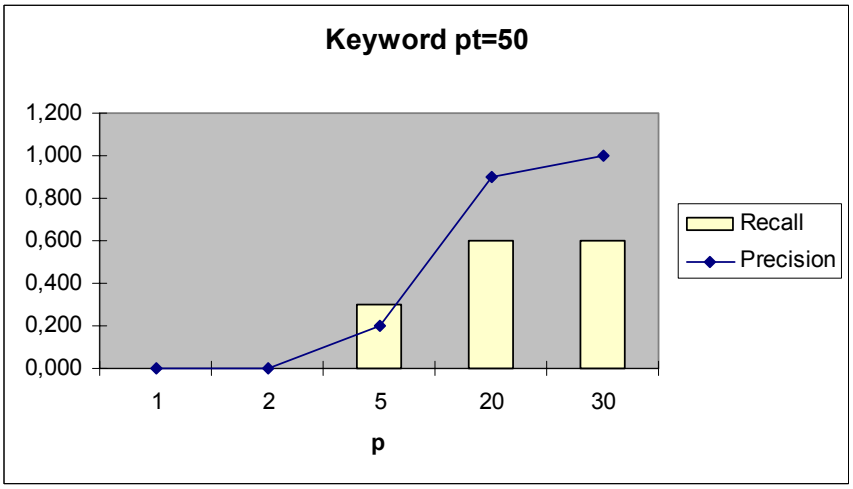
5.2.3 Υπολογισμός precision-recall για την εκτίμηση των προτιμήσεων

Σε αυτή την παράγραφο χρησιμοποιήσαμε μια επιπλέον μετρική για την αξιολόγηση των παραπάνω μεθοδολογιών. Συγκεκριμένα χρησιμοποιήσαμε την απλή εξίσωση υπολογισμού των προτιμήσεων (TUA + (FUA - TUA) * (1 - $\frac{1}{\left[\frac{n/M}{k/N}\right]^2 + 2}$)). Η διαδικασία ήταν η ίδια με αυτήν που ακολουθήθηκε στις προηγούμενες περιπτώσεις. Μετά την εκτίμηση των προτιμήσεων για τα προγράμματα του test set επιλέξαμε από αυτά εκείνα με τις καλύτερες βαθμολογίες χρησιμοποιώντας κάποιο κατώφλι. Δοκιμάσαμε διάφορα κατώφλια και κρατήσαμε αυτό που μας έδωσε τις καλύτερες τιμές. Στη συνέχεια από τις πραγματικές βαθμολογίες των χρηστών κρατήσαμε αυτές που ήταν πάνω από 4. Τέλος για τα δύο αυτά σύνολα υπολογίσαμε το precision και recall προκειμένου να έχουμε μια ένδειξη της επιτυχίας της μεθόδου για την εκτίμηση των πιο ενδιαφερόντων για κάθε χρήστη προγραμμάτων. Τα αποτελέσματα που πήραμε φαίνονται στα παρακάτω διαγράμματα:



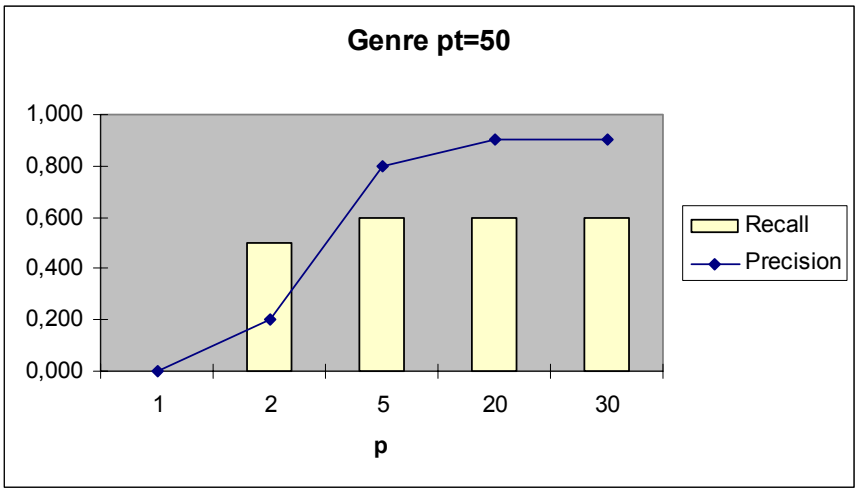
Διάγραμμα 19: Precision και Recall για προφίλ με χρήση του actor. Υπολογίζεται η προτίμηση του χρήστη για τα προγράμματα test set με βάση τα χαρακτηριστικά του train set και τον τύπο(TUA-(FUA-TUA)*(1-1/r^2+2)). Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=50), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: Παρατηρούμε πολύ καλή επίδοση για το precision για μεγάλες τιμές του p ενώ το recall κρατιέται σε πιο χαμηλά επίπεδα που θεωρούνται όμως αρκετά ικανοποιητικά.



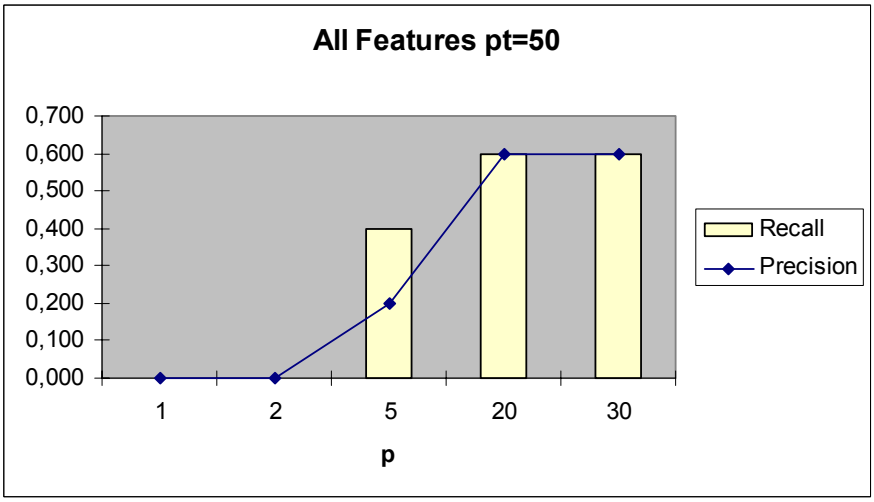
Διάγραμμα 20: Precision και Recall για προφίλ με χρήση του keyword. Υπολογίζεται η προτίμηση του χρήστη για τα προγράμματα test set με βάση τα χαρακτηριστικά του train set και τον τύπο(TUA-(FUA-TUA)*(1-1/r^2+2)). Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=50), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: έχουμε πολύ καλή επίδοση για το precision, καλύτερη από αυτή που είχαμε για τους actors ενώ το recall και πάλι βρίσκεται σε χαμηλότερες τιμές και πάλι όμως ικανοποιητικές.



Διάγραμμα 21: Precision και Recall για προφίλ με χρήση του genre. Υπολογίζεται η προτίμηση του χρήστη για τα προγράμματα test set με βάση τα χαρακτηριστικά του train set και τον τύπο(TUA-(FUA-TUA)*(1-1/r^2+2)). Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=50), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: η τιμές για το recall είναι ακόμα πιο υψηλές από τις περιπτώσεις των προηγούμενων χαρακτηριστικών ενώ υπάρχει βελτίωση του recall και για μικρότερες τιμές τις παραμέτρου p.



Διάγραμμα 22: Precision και Recall για προφίλ με χρήση του genre. Υπολογίζεται η προτίμηση του χρήστη για τα προγράμματα test set με βάση τα χαρακτηριστικά του train set και τον τύπο(TUA-(FUA-TUA)*(1-1/r^2+2)). Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=50), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: παρατηρούμε πως για μικρές τιμές του p έχουμε κακή επίδοση και των δύο μεγεθών ενώ για μεγάλα p παρουσιάζονται ίδιες επιδόσεις για το recall και το precision.

Συμπεράσματα

Για το precision πήραμε πολύ καλές τιμές για όλα τα χαρακτηριστικά με καλύτερο την κατηγορία των προγραμμάτων. Για το recall οι καλύτερες τιμές ήταν ίδιες για όλες τις περιπτώσεις. Τέλος για το συνδυασμό των χαρακτηριστικών είχαμε εξίσου καλό recall με τα μεμονωμένα χαρακτηριστικά. Ανάλογες ήταν οι τιμές για το precision, τιμές που κινήθηκαν όμως πιο χαμηλά από αυτές που είχαμε για την περίπτωση της κατηγορίας των προγραμμάτων.

5.2.4 Πείραμα: χρήση μόνο των χαρακτηριστικών των προγραμμάτων

Μέχρι σε αυτό το σημείο η προσέγγιση που κάναμε προσπάθησε να λάβει υπόψη της τόσο την συμπεριφορά των χαρακτηριστικών των προγραμμάτων, μελετώντας την συχνότητα εμφάνισής τους, όσο και την συμπεριφορά των βαθμολογιών του χρήστη χρησιμοποιώντας μέτρα όπως αυτό της εντροπίας. Στη τελευταία ομάδα που θα παρουσιάσουμε, προσπαθήσαμε να κάνουμε εκτίμηση των προτιμήσεων των χρηστών βασισμένοι μόνο στα χαρακτηριστικά που εμφανίζονται στα αξιολογημένα προγράμματα.

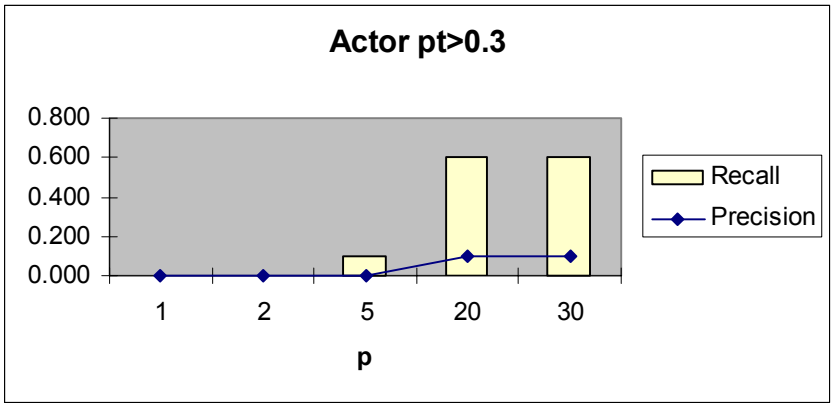
Συγκεκριμένα για κάθε χρήστη χωρίσαμε τα δεδομένα πρώτα σε 80% train data και 20% test data και στη συνέχεια σε 50-50% αντίστοιχα. Από τα train data κρατήσαμε τα προγράμματα με τις καλύτερες βαθμολογίες(4-5 στην αντίστοιχη κλίμακα) και με βάση αυτά προσπαθήσαμε να υπολογίσουμε την πιθανότητα να δει καθένα από τα προγράμματα του test set. Αναλυτικότερα, για κάθε τιμή χαρακτηριστικού f_i των προγραμμάτων υπολογίσαμε την πιθανότητα να δει ο χρήστης το συγκεκριμένο πρόγραμμα δοθέντος ότι περιέχει το αντίστοιχο χαρακτηριστικό. Με αυτόν τον τρόπο καταλήξαμε σε μια τιμή πιθανότητας για κάθε μεμονωμένο χαρακτηριστικό. Ο αντίστοιχος τύπος υπολογισμού ήταν: $P(see \mid f_i) = \frac{P(f_i \mid seen)P(seen)}{P(f_i \mid seen)P(seen) + P(f_i \mid notseen)P(notseen)} =$

$$\frac{n_i^u / n^u * n^u / n^s}{\frac{n_i^u}{n^u} * \frac{n^u}{n^s} + \frac{n_i^s - n_i^u}{n^s - n^u} * \frac{n^s - n^u}{n^s}} = \frac{n_i^u}{n_i^s}$$

όπου n^s, n^u το πλήθος των προγραμμάτων

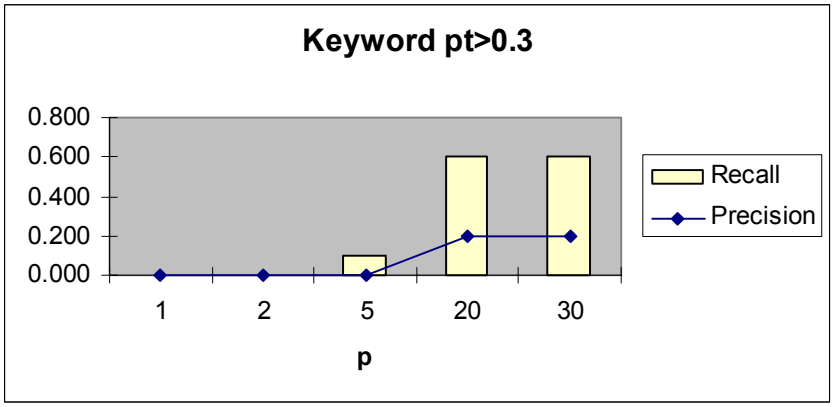
του συστήματος και του χρήστη αντίστοιχα και n_i^u , n_i^s το πλήθος των προγραμμάτων του χρήστη και του συστήματος αντίστοιχα που περιέχουν το χαρακτηριστικό f_i.

Αφού έγινε ο υπολογισμός των πιθανοτήτων για όλα τα χαρακτηριστικά ακολούθησε ο συνδυασμός τους σε ένα FASP. Η συσχέτισή του με τα προγράμματα του test set μας επέστρεψε την πιθανότητα να θέλει να δει ο χρήστης καθένα από τα προγράμματα. Στη συνέχεια έγινε ταξινόμηση των προγραμμάτων κάθε χρήστη με βάση την πιθανότητά τους και κρατήθηκαν αυτά που βρέθηκαν πάνω από κάποιο κατώφλι που ορίσαμε. Τέλος με βάση τις πραγματικές τιμές αξιολόγησης των προγραμμάτων κρατήσαμε τα αντίστοιχα με τις υψηλότερες βαθμολογίες. Για τα δύο σύνολα και για κάθε περίπτωση υπολογίσαμε το precision και το recall. Τα αποτελέσματα φαίνονται αναλυτικά στα παρακάτω διαγράμματα:



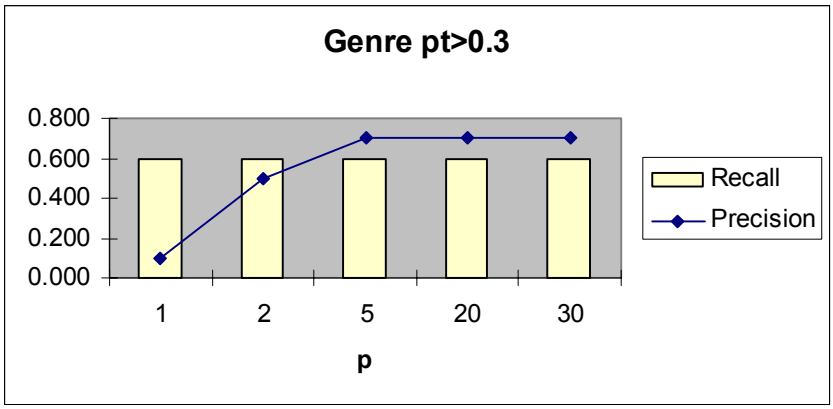
Διάγραμμα 23: Precision και Recall για προφίλ με χρήση του actor. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: παρατηρούμε πως για μικρές τιμές του p δεν έχουμε καμιά σωστή πρόβλεψη και είναι 0 τόσο το precision όσο και το recall. Για μεγάλες τιμές έχω σχετικά καλό recall αλλά το precision παραμένει σε πολύ μικρές τιμές.



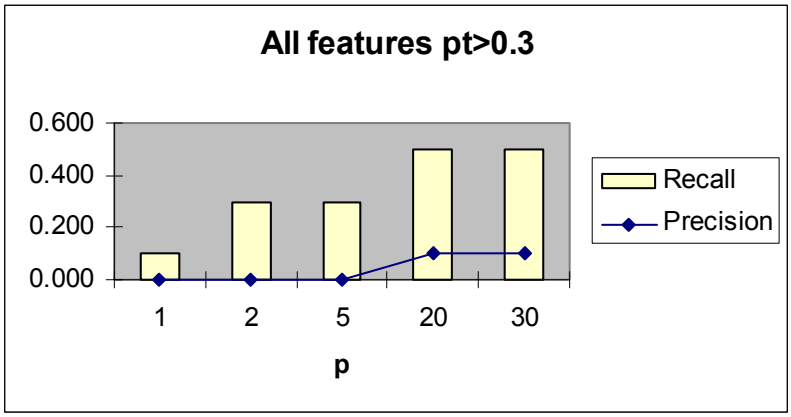
Διάγραμμα 24: Precision και Recall για προφίλ μόνο μ ε το keyword. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: η συμπεριφορά για το keyword παρουσιάζεται ίδια με αυτήν που είδαμε προηγουμένως για τον actor. Υπάρχει ελαφρώς καλύτερο precision αλλά παραμένει σε πολύ χαμηλές τιμές.



Διάγραμμα 25: Precision και Recall για προφίλ με χρήση του genre. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: Στη περίπτωση του genre έχουμε πολύ καλές επιδόσεις για όλες τις τιμές του p για το recall ενώ το precision παρουσιάζει μια βελτίωση και τελικά σταθεροποίηση καθώς αυξάνει το p. Οι τιμές του recall φτάνουν το 0.6 ενώ του precision υπερβαίνει το 0.7, τιμές πολύ ικανοποιητικές.



Διάγραμμα 26: Precision και Recall για προφίλ με συνδυασμό των χαρακτηριστικών. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: για το συνδυασμό των χαρακτηριστικών έχουμε παρόμοια συμπεριφορά με αυτή που είχαμε στο keyword και στον actor, όμως υπάρχει κάποιο χαμηλό recall για τα μικρά p.

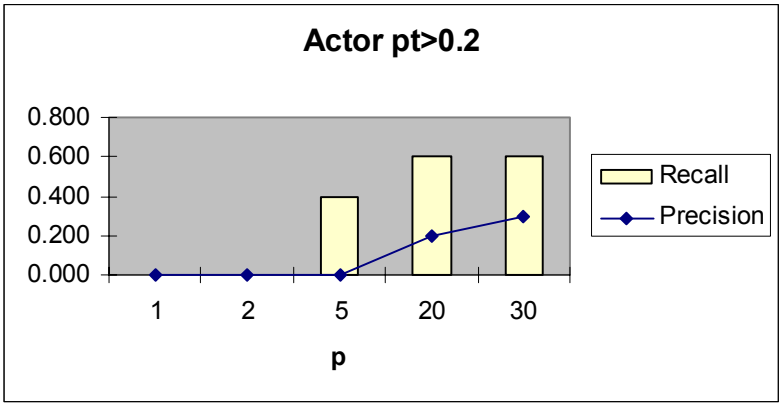
Συμπεράσματα

Το πιο σημαντικό που ίσως έχουμε να σημειώσουμε είναι η πολύ καλή συμπεριφορά για τα genres. Είναι πολύ καλύτερη από τα άλλα δύο χαρακτηριστικά ενώ παρουσιάζεται καλύτερη και από το συνδυασμό των χαρακτηριστικών. Το πρώτο θα μπορούσαμε να το αποδώσουμε στη φύση των δεδομένων που διαθέτουμε. Συγκεκριμένα τα χαρακτηριστικά που διαθέτουνε είναι τέτοια ώστε ο ηθοποιός και οι λέξεις κλειδιά να είναι σύνολα με πολύ μεγάλη πληθυκότητα και με λίγες τιμές που να επαναλαμβάνονται στα προγράμματα. Αντίθετα οι κατηγορίες των προγραμμάτων είναι πολύ λίγες και επαναλαμβάνονται πολύ συχνά οι περισσότερες. Στο πείραμα που κάναμε επιλέξαμε μόνο τα προγράμματα με τις καλύτερες βαθμολογίες. Αυτό πρακτικά σημαίνει πως οι ηθοποιοί και οι λέξεις κλειδιά που περιέχονται σε αυτά τα λίγα προγράμματα του train set έχουν πολύ μικρότερη πιθανότητα σε σχέση με τις κατηγορίες των προγραμμάτων να εμφανίζονται στα προγράμματα του test set. Σύμφωνα με τα παραπάνω θα μπορούσε να

δικαιολογηθεί η κακή επίδοση των δύο πρώτων χαρακτηριστικών σε σχέση με την κατηγορία των προγραμμάτων.

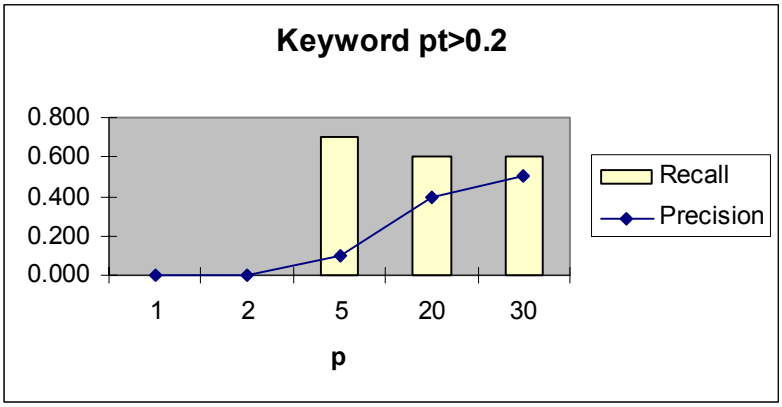
Εκείνο που πιθανώς δεν θα ανέμενε κανείς είναι η κακή επίδοση του συνδυασμού των χαρακτηριστικών. Παρ' όλα αυτά θα πρέπει να θυμηθούμε πως με βάση το TVA όλα τα χαρακτηριστικά διαφορετικών κατηγοριών συνδυάζονται με το ίδιο βάρος μεταξύ τους. Αυτό το γεγονός δεν μας επιτρέπει κατά τον συνδυασμό των χαρακτηριστικών να δώσουμε ένα μεγαλύτερο βάρος ενδεχομένως στη κατηγορία των προγραμμάτων που έχει καλύτερη συμπεριφορά σε σχέση με τα άλλα χαρακτηριστικά. Αυτό έχει σαν αποτέλεσμα ο συνδυασμός των χαρακτηριστικών να παρουσιάζει την κακή απόδοση που είδαμε παραπάνω επηρεασμένος από την ανάλογη απόδοση των δύο από τα τρία χαρακτηριστικά που περιέχει.

Στη συνέχεια θα παρουσιάσουμε τις καμπύλες για το precision και το recall για τον ίδιο χωρισμό από δεδομένα. Η διαφορά είναι πως σε αυτή την περίπτωση το κατώφλι επιλογής(pt) των καλύτερων προγραμμάτων από τις προβλέψεις είναι 0.2 αντί 0.3 που ήταν πριν.



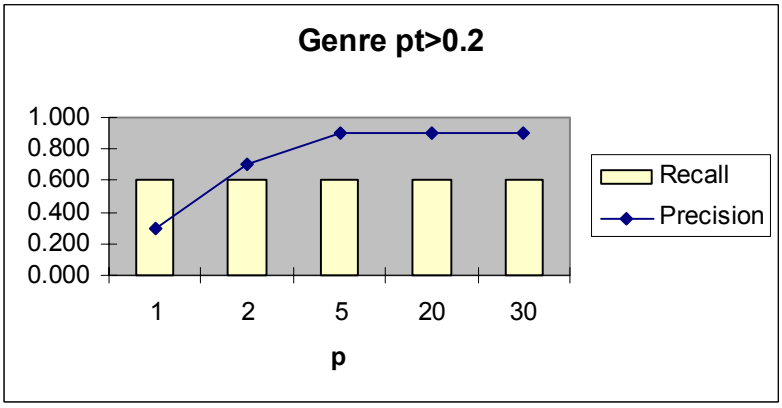
Διάγραμμα 27: Precision και Recall για προφίλ με μόνο μ ε τον Actor. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: σε σχέση με την προηγούμενη ομάδα πειραμάτων και το αντίστοιχο διάγραμμα(23), παρατηρούμε μια βελτίωση για το precision καθώς μεγαλώνει το p.



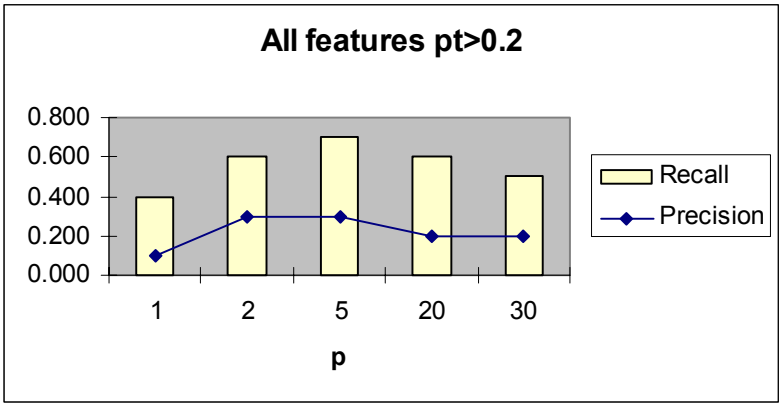
Διάγραμμα 28: Precision και Recall για προφίλ με μόνο μ ε το Keyword. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: σε σχέση με το αντίστοιχο διάγραμμα(24) παρατηρούμε βελτίωση τόσο για το recall όσο και αρκετά σημαντική βελτίωση για το precision.



Διάγραμμα 29: Precision και Recall για προφίλ με μόνο μ ε το Genre. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

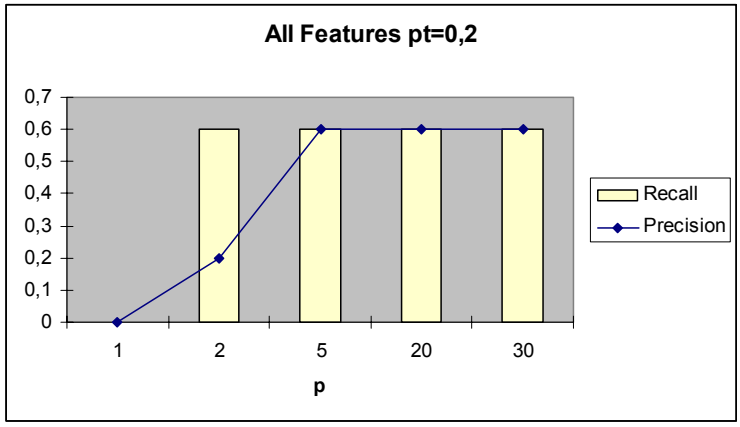
Παρατηρήσεις: και σε αυτή την περίπτωση το genre παρουσιάζει πολύ καλές επιδόσεις τόσο για το precision όσο και για το recall. Το recall μάλιστα φαίνεται να είναι ελαφρώς βελτιωμένο.



Διάγραμμα 30: Precision και Recall για προφίλ με το συνδυασμό των χαρακτηριστικών. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: παρατηρούμε σχετικά καλές τιμές για το recall ενώ και precision είναι βελτιωμένο σε σχέση με αυτό του διαγράμματος 26. Παρ' όλα αυτά παραμένει σε χαμηλές τιμές.

Στο παραπάνω διάγραμμα έγινε εφαρμογή της μεθοδολογίας χρησιμοποιώντας ίδια βάρη για όλα τα χαρακτηριστικά. Για αυτό και παρατηρήσαμε αυτή την κακή συμπεριφορά σε σχέση με το genre. Θα πρέπει βέβαια να σημειώσουμε πως αυτή είναι και η μόνη δυνατότητα κατά το TVA. Παρ' όλα αυτά στην υλοποίηση υποστηρίξαμε την δυνατότητα χρήσης βαρών μεταξύ διαφορετικών χαρακτηριστικών κατά τη διαδικασία της συσχέτισης. Με αυτό τον τρόπο τα αποτελέσματα που πήραμε για το συνδυασμό των χαρακτηριστικών ήταν τα εξής:



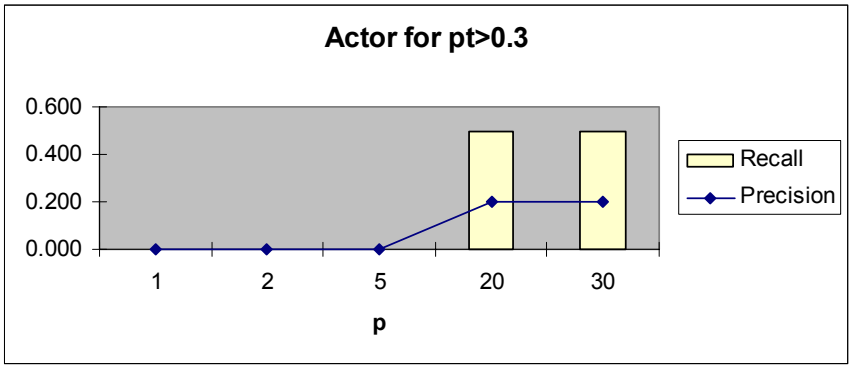
Διάγραμμα 31: Precision και Recall για προφίλ με το συνδυασμό των χαρακτηριστικών. Χρησιμοποιήθηκαν βάρη με σ΄χεση 1/0.4 για το συνδυασμό των genre με τα άλλα δύο χαρακτηριστικά Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 50% train data, 50% test data.

Παρατηρήσεις: όπως φαίνεται υπάρχει πολύ σημαντική βελτίωση σε σχέση με τα αποτελέσματα που είχαμε πάρει στο προηγούμενο διάγραμμα(30) όπου όλα τα χαρακτηριστικά είχαν συνδυαστεί με το ίδιο βάρος. Με αυτόν το τρόπο καταφέραμε να πάρουμε αποτελέσματα πολύ κοντά σε αυτά που πήραμε με το genre που ήταν και το καλύτερο από τα άλλα δύο.

Συμπεράσματα

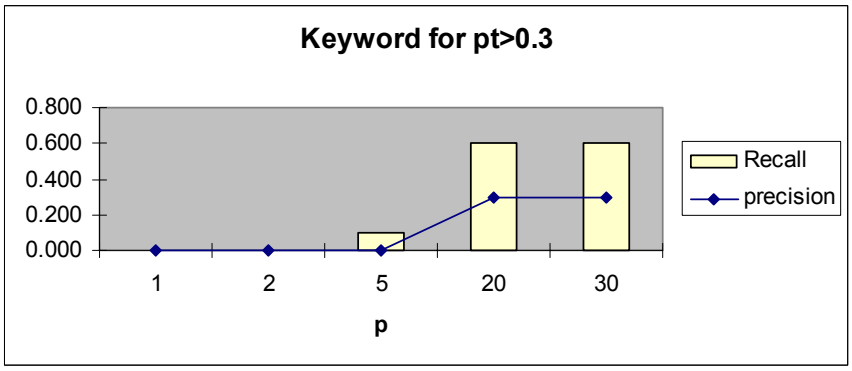
Αυτό που θα πρέπει να σημειώσουμε είναι τα καλύτερα αποτελέσματα που πήραμε για όλες τις περιπτώσεις σε σχέση με αυτά που είχαμε για p=0.3. Αυτό διαπιστώσαμε και πρακτικά πως οφείλετε στο γεγονός ότι η διαδικασία συσχέτισης του προφίλ, με τις πιθανότητες, με τα προγράμματα του test set, επέστρεψε γενικώς μικρές τιμές πιθανοτήτων ως πρόβλεψη. Η μέση του τιμή σε πολλές περιπτώσεις κινείται κάτω από το 0.2. Επομένως είναι λογικό να απαιτείται μικρό κατώφλι για την επιλογή των καλύτερων προγραμμάτων. Τέλος για το συνδυασμό των χαρακτηριστικών, η βελτίωση των τιμών οφείλεται στη βελτίωση των επιδόσεων των επί μέρους χαρακτηριστικών. Παρ' όλα αυτά παραμένουν χειρότερες από αυτές που είχαμε για την περίπτωση της κατηγορίας των προγραμμάτων.

Ακολουθούν δύο ακόμα ομάδες προγραμμάτων για κατώφλια 0.3 και 0.2 όπως στις προηγούμενες περιπτώσεις. Η διαφορά τώρα είναι πως χωρισμός των δεδομένων ήταν 80%-20% σε train set και test set αντίστοιχα.



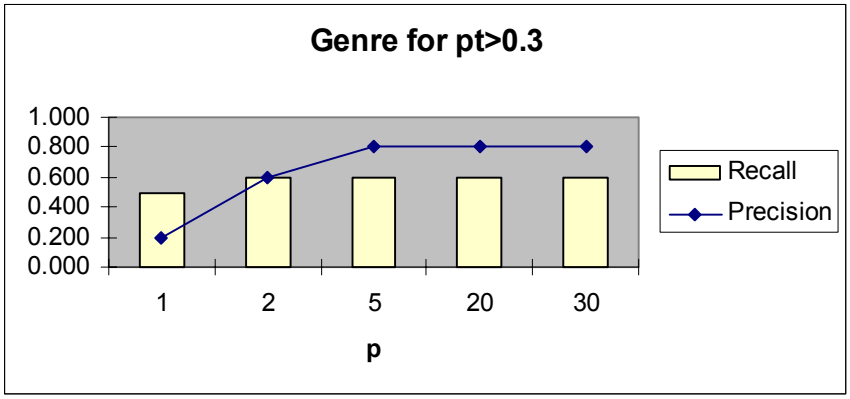
Διάγραμμα 32: Precision και Recall για προφίλ με χρήση του actor. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: η τιμές που πήραμε ήταν ανάλογες αυτών που παρατηρήθηκαν για το 50-50 χωρισμό δεδομένων στο διάγραμμα(23). Υπάρχει μια χειροτέρευση για το recall.



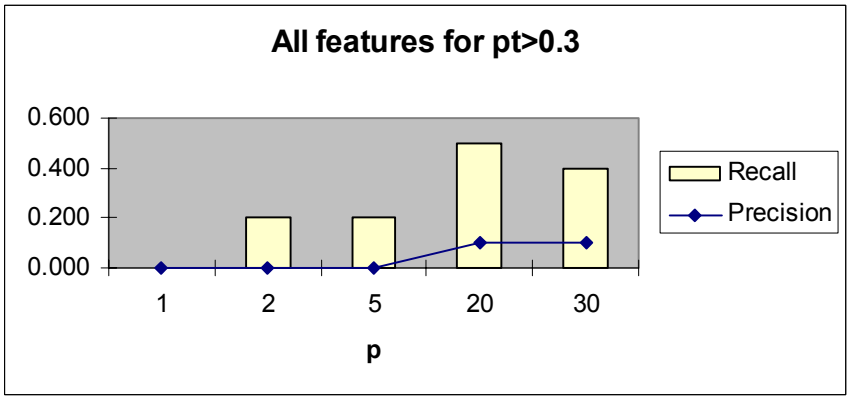
Διάγραμμα 33: Precision και Recall για προφίλ με χρήση του Keyword. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: και σε αυτή την περίπτωση η συμπεριφορά είναι ανάλογη αυτής του διαγράμματος 24. παρατηρούμε πως το precision για τις περιπτώσεις του p=20, 30 είναι βελτιωμένο.



Διάγραμμα 34: Precision και Recall για προφίλ με χρήση του Genre. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: το genre συνεχίζει να έχει την καλύτερη συμπεριφορά. Η διαφορά με το διάγραμμα 25 είναι πως έχουμε καλύτερο precision και για την τιμή 2 του p.



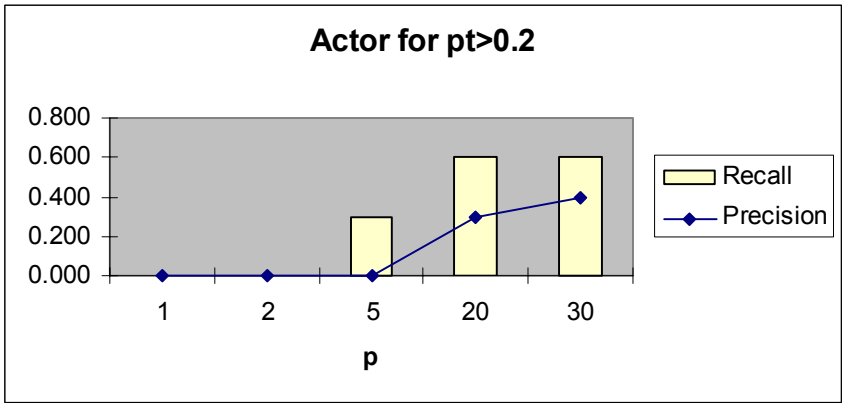
Διάγραμμα 35: Precision και Recall για προφίλ με το συνδυασμό των χαρακτηριστικών. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.3), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: Παρατηρούμε πως ο συνδυασμός των χαρακτηριστικών δεν μας δίνει και πάλι καλές τιμές και η συμπεριφορά είναι ανάλογη αυτής που είχαμε στην περίπτωση του διαγράμματος 26.

Συμπεράσματα

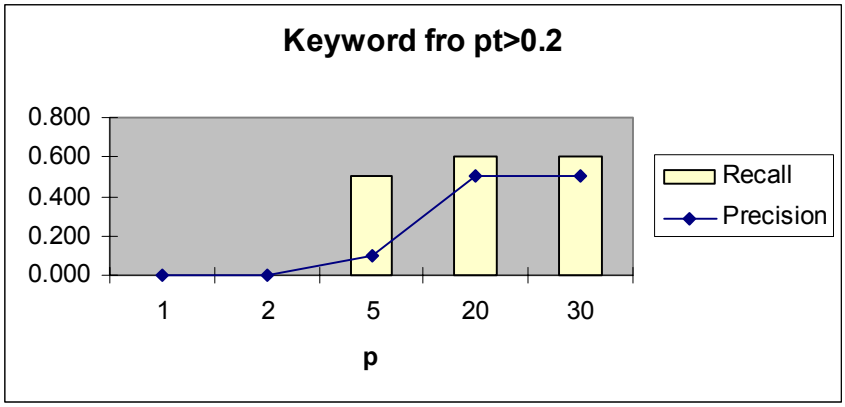
Σε αυτήν την ομάδα των πειραμάτων η συμπεριφορά του συστήματος ήταν ανάλογη αυτή που είχαμε για χωρισμό δεδομένων σε 50-50% με μόνο κάποια βελτίωση σε περιορισμένες τιμές κυρίως του precision.

Η ομάδα πειραμάτων που παρουσιάζουμε στη συνέχεια είναι πάλι για χωρισμό δεδομένων 80-20% αλλά αυτή τη φορά κατώφλι επιλογής το 0.2.



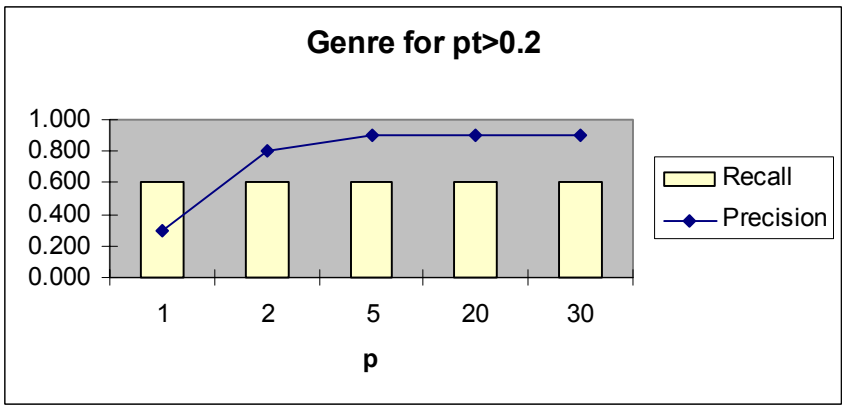
Διάγραμμα 36: Precision και Recall για προφίλ μόνο μ ε τον Actor. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: έχουμε μια βελτίωση αντίστοιχης αυτής που παρατηρήσαμε πάλι για κατώφλι 0.2 στο διάγραμμα 28.



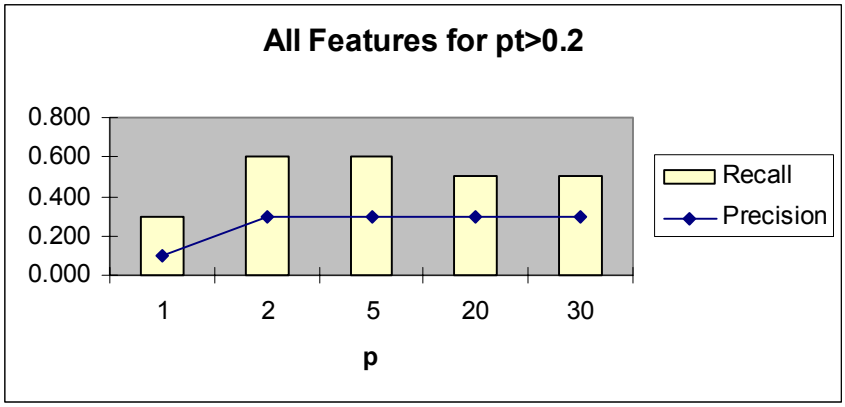
Διάγραμμα 37: Precision και Recall για προφίλ μόνο μ ε το Keyword. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: και στη περίπτωση του keyword παρατηρούμε πολύ σημαντική βελτίωση, κυρίως για το precision, για κατώφλι 0.2.



Διάγραμμα 38: Precision και Recall για προφίλ μόνο μ ε το Genre. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

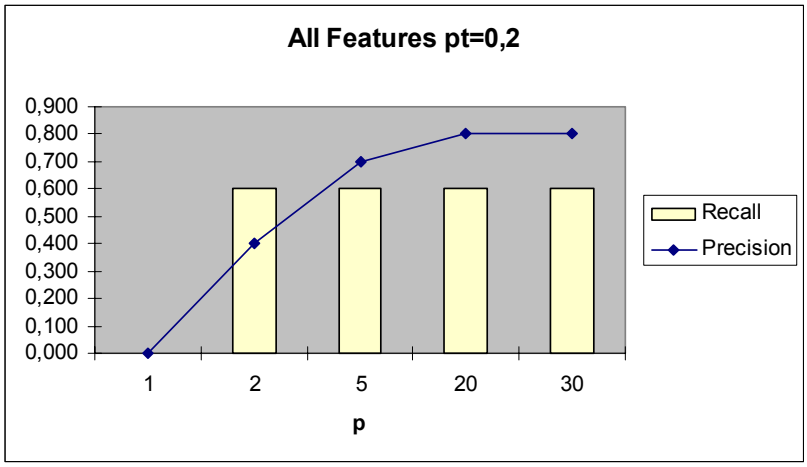
Παρατηρήσεις: σε αυτήν τη περίπτωση έχουμε μια ακόμα μεγαλύτερη βελτίωση από αυτήν που είχαμε στην αντίστοιχη περίπτωση του διαγράμματος 29.



Διάγραμμα 39: Precision και Recall για προφίλ μόνο μ ε συνδυασμό των χαρακτηριστικών. Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: σε αυτή τη περίπτωση παρατηρούμε σχετικά σταθερή συμπεριφορά για το precision και βελτιωμένη σε σχέση με την αντίστοιχη περίπτωση του διαγράμματος 30.

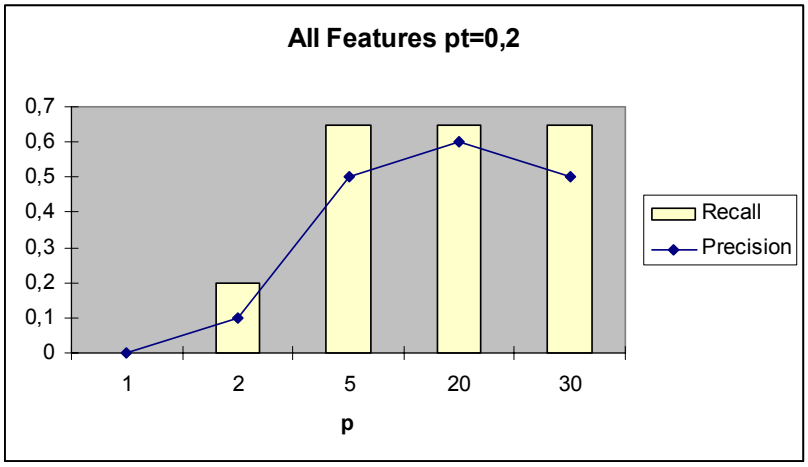
Και σε αυτήν τη περίπτωση ο συνδυασμός των χαρακτηριστικών έγινε χρησιμοποιώντας ίδια βάρη για όλα τα χαρακτηριστικά. Έτσι η συμπεριφορά επηρεάστηκε από την συμπεριφορά του actor και του keyword και δεν μπορέσαμε να προσεγγίσουμε την καλή συμπεριφορά του genre. Στη συνέχεια χρησιμοποιήσαμε βάρη για το συνδυασμό του genre με τα υπόλοιπα χαρακτηριστικά δίνοντάς του πολύ μεγαλύτερο βάρος και τα αποτελέσματα που πήραμε ήταν τα εξής:



Διάγραμμα 40: Precision και Recall για προφίλ μόνο μ ε συνδυασμό των χαρακτηριστικών. Χρησιμοποιήθηκαν βάρη με σχέση 1/0.4 για το συνδυασμό του genre με τα άλλα δύο χαρακτηριστικά Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

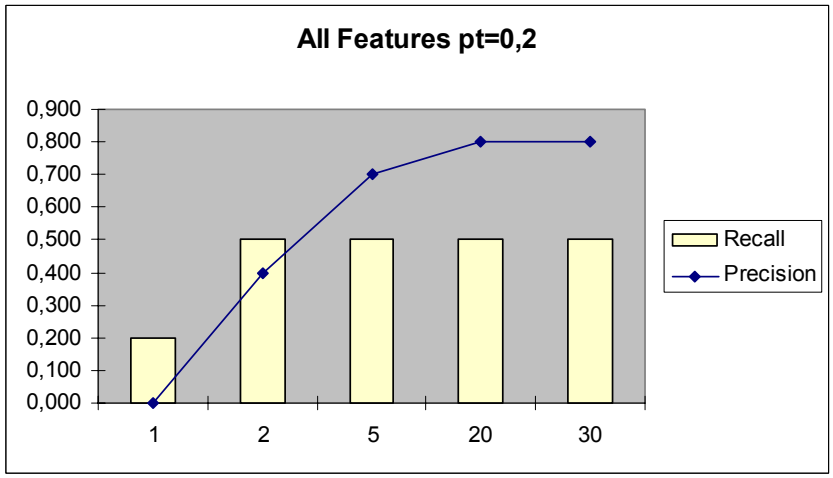
Παρατηρήσεις: τα αποτελέσματα που πήραμε δίνοντας κατά τη συσχέτιση πολύ μεγαλύτερο βάρος στο χαρακτηριστικό με την καλύτερη συμπεριφορά(genre) ήταν πολύ βελτιωμένα και προσέγγισαν την συμπεριφορά του καθώς η επίδραση των άλλων δύο ήταν πολύ μικρότερη.

Στα επόμενα δύο διαγράμματα παρουσιάζουμε τα αποτελέσματα που πήραμε για βάρος 0.8 και 0.2 αντίστοιχα 0.4. Το βάρος για το genre παρέμεινε 1.



Διάγραμμα 41: Precision και Recall για προφίλ μόνο μ ε συνδυασμό των χαρακτηριστικών. Χρησιμοποιήθηκαν βάρη με σχέση 1/0.8 για το συνδυασμό του genre με τα άλλα δύο χαρακτηριστικά Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: παρατηρούμε πως καθώς ενισχύθηκε η επίδραση των actor και keyword είχαμε χειρότερες τιμές κυρίως για το precision.



Διάγραμμα 42: Precision και Recall για προφίλ μόνο μ ε συνδυασμό των χαρακτηριστικών. Χρησιμοποιήθηκαν βάρη με σχέση 1/0.2 για το συνδυασμό του genre με τα άλλα δύο χαρακτηριστικά Υπολογίζεται η πιθανότητα να δει ο χρήστης κάθε πρόγραμμα του test set με βάση τα χαρακτηριστικά του train set. Το pt είναι κατώφλι επιλογής για το test set. **Παράμετροι:** κατώφλι επιλογής το pt(=0.2), διαφορετικά p κατά τη συσχέτιση, χωρισμός δεδομένων 80% train data, 20% test data.

Παρατηρήσεις: σε αυτή τη περίπτωση μειώσαμε ακόμα περισσότερο το βάρος για τα actor και keyword. Παρατηρείται μικρή βελτίωση κυρίως για το recall.

Συμπεράσματα

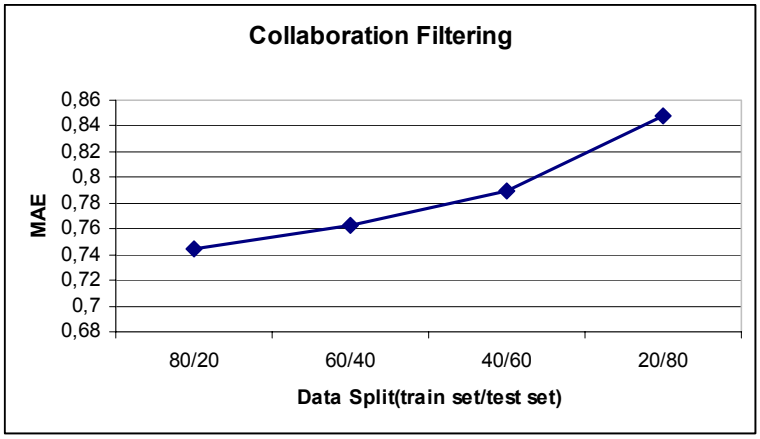
Και σε αυτή την ομάδα πειραμάτων, η αλλαγή του χωρισμού των δεδομένων δεν άλλαξε σημαντικά τη συμπεριφορά του συστήματος. Η κατηγορία των προγραμμάτων παρέμεινε η περίπτωση με τις καλύτερες επιδόσεις. Τέλος η σχετική συμπεριφορά με τα υπόλοιπα χαρακτηριστικά παρέμεινε ίδια τόσο για τους δυο διαφορετικούς χωρισμούς δεδομένων που δοκιμάσαμε όσο και για τα διαφορετικά κατώφλια που επιλέξαμε.

5.2.5 Συγκριτική παρουσίαση συνεργατικού φιλτραρίσματος.

Στα πλαίσια των πειραμάτων που διεξήχθησαν χρησιμοποιήσαμε και παραδοσιακές μεθόδους συνεργατικού φιλτραρίσματος προκειμένου να έχουμε ένα μέτρο σύγκρισης για τις επιδόσεις του συστήματός μας. Συγκεκριμένα χρησιμοποιήσαμε την ομοιότητα μεταξύ χρηστών για να κάνουμε προβλέψεις με τη βοήθεια της εξίσωσης του Pearson. Αναλυτικότερα, έγινε χρήση μόνο των αξιολογήσεων των χρηστών για τα προγράμματα του train set προκειμένου να γίνει πρόβλεψη για τα προγράμματα του test set. Χρησιμοποιήσαμε και πάλι τέσσερις διαφορετικούς χωρισμούς δεδομένων. Τέλος θα θυμίσουμε την εξίσωση pearson η οποία είναι η εξής:

$$s_{u,v} = \frac{\sum_{i \in \mathrm{A}} \left(R_{u,i} - \overline{R}_u\right) \cdot \left(R_{v,i} - \overline{R}_v\right)}{\sqrt{\sum_{i \in A} \left(R_{u,i} - \overline{R}_u\right)^2} \cdot \sqrt{\sum_{i \in A} \left(R_{v,i} - \overline{R}_v\right)^2}}$$

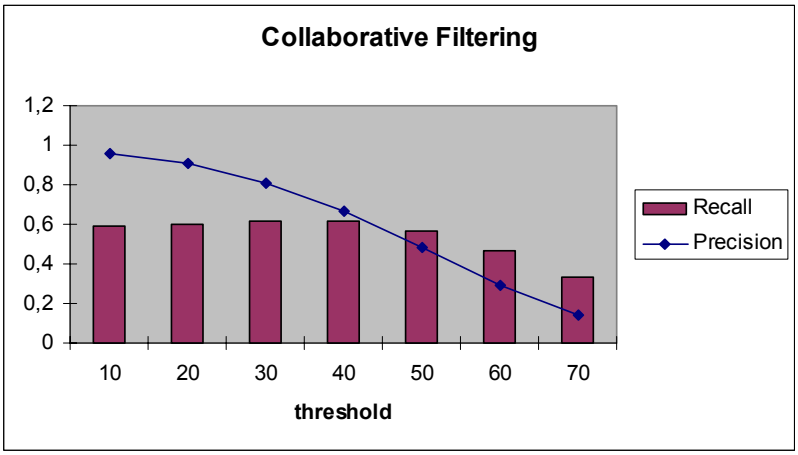
Αυτός ο τύπος υπολογίζει την ομοιότητα $s_{u,v}$ κάθε ζεύγους χρηστών *u* και *v*. Α είναι ένα σύνολο προγραμμάτων βαθμολογημένο και από τους δύο χρήστες. Στο επόμενο διάγραμμα φαίνονται οι τιμές που πήραμε για το MAE για τους τέσσερις διαφορετικούς χωρισμούς από δεδομένα που χρησιμοποιήθηκαν:



Διάγραμμα 43: MAE για τα δεδομένα των πειραμάτων με χρήση user-to-user ομοιότητα.

Παρατηρήσεις: παρατηρούμε πως για τις καλύτερες περιπτώσεις του συνεργατικού φιλτραρίσματος το MAE είναι αρκετά μικρότερο από αυτό που επιτύχαμε με τις μεθοδολογίες που παρουσιάσαμε παραπάνω αφού το μικρότερο που επιτύχαμε ήταν περίπου 0.840. Στις παρατηρήσεις και τα συμπεράσματά μας, έχουμε εντοπίσει αδυναμίες του συστήματος ενώ επιπλέον δουλειά χρειάζεται και σε αυτήν αναφερόμαστε στο επόμενο κεφάλαιο. Τέλος θα πρέπει να θυμίσουμε τις απαιτήσεις που έχουν τεχνικές όπως αυτή του συνεργατικού φιλτραρίσματος. Αντίθετα ένα σύστημα σαν αυτό που αναπτύξαμε παρουσιάζει πολύ καλύτερες χρονικές αποδόσεις ενώ και οι απαιτήσεις του είναι αρκετά πιο περιορισμένες.

Τέλος στο επόμενο διάγραμμα παρουσιάζουμε τις τιμές για τα precision, recall που πήραμε με χρήση του collaborative filtering. Συγκεκριμένα από τις προβλέψεις που πήραμε κρατήσαμε τις καλύτερες με βάση κάποιο κατώφλι και τις συγκρίναμε με αυτές από τις πραγματικές βαθμολογίες που είχανε βαθμό πάνω από 4. Οι τιμές για τα διάφορα κατώφλια ήταν οι εξής:



Διάγραμμα 44: Precision και Recall για τις προβλέψεις που πήραμε από το collaborative filtering. Ο χωρισμός των δεδομένων ήταν 80-20% και το threshold ήταν κατώφλι επιλογής για τα προγράμματα με βάση τις τιμές των εκτιμήσεων.

Παρατηρήσεις: η συμπεριφορά του recall είναι σχετικά σταθερή ενώ το precision μειώνεται καθώς μειώνεται το σύνολο των προγραμμάτων που επιλέγουμε με το κατώφλι. Εδώ θα πρέπει να σημειώσουμε πως η συμπεριφορά του genre ήταν εξίσου καλή και κατά περιπτώσεις με βελτιωμένο και το recall ενώ το precision παρέμενε καλό. Τέλος θα πρέπει να σημειώσουμε πως το recall είχε τις καλύτερες τιμές για κατώφλι 30 και 40. Τόσο για μεγαλύτερο όσο και για μικρότερο κατώφλι αρχίζει να σημειώνει πτώση.

5.2.6 Συνολικά Συμπεράσματα Πειραμάτων

Σε αυτό το κεφάλαιο μελετήσαμε την απόδοση των διαφόρων μεθοδολογιών που αναπτύχθηκαν. Συνοπτικά εφαρμόσαμε απλές τεχνικές εκτίμησης των προτιμήσεων των χρηστών για τα διάφορα χαρακτηριστικά. Οι αντίστοιχη τύποι βρίσκονται συγκεντρωμένοι την παράγραφο 5.2.2. Από τα αποτελέσματα που πήραμε θα μπορούσαμε κατ' αρχή να παρατηρήσουμε πως με την χρήση καμπύλης σύγκλισης από τον γενικό μέσω όρο των βαθμολογιών του χρήστη στο μέσο όρο των βαθμολογιών για κάθε χαρακτηριστικό, είχαμε σημαντική βελτίωση των αποτελεσμάτων με βάση το MAE που πήραμε για αυτήν την

περίπτωση και την περίπτωση χρήσης απλά του μέσου όρου των προγραμμάτων για κάθε χαρακτηριστικό. Αντίθετα η χρήση της εντροπίας και της διασποράς των βαθμολογιών του χρήστη δεν μας έδωσε τα αναμενόμενα αποτελέσματα που αφού δεν είχαμε περαιτέρω βελτίωση της απόδοσης του συστήματος. Στη συνέχεια δοκιμάζοντας να εκτιμήσουμε τις προτιμήσεις των χρηστών βασισμένοι μόνο στα χαρακτηριστικά των προγραμμάτων πήραμε πολύ ικανοποιητικές τιμές για το precision και το recall με βάση την ικανότητα του συστήματος να εκτιμήσει τα προγράμματα με την μεγαλύτερη για το χρήστη προτίμηση. Παράλληλα τα αποτελέσματα αυτά ήταν ανάλογα των αποτελεσμάτων που πήραμε και για τις προηγούμενες τεχνικές. Τέλος εφαρμόσαμε τις ίδιες μετρικές για συνεργατικό φιλτράρισμα μεταξύ χρηστών και τα αποτελέσματα που πήραμε ήταν αρκετά καλύτερα για το MAE ενώ για το precision και το recall ήταν παρόμοια με του δικού μας συστήματος. Σε αυτό το σημείο θα πρέπει να λάβουμε υπόψη τις απαιτήσεις από το σύστημα, τεχνικών όπως το συνεργατικό φιλτράρισμα οι οποίες είναι πολύ μεγάλες και σε ένα σύστημα σαν το up-TV θα μπορούσαν να τρέχουν μόνο από τη πλευρά κάποιου κεντρικού εξυπηρετητή. Αντίθετα το σύστημα που αναπτύξαμε έχει πολύ μικρότερες απαιτήσεις και στην περίπτωση που παρουσιάζει εφ' άμιλλες επιδόσεις θα μπορούσε να χρησιμοποιείται τόσο στη πλευρά του server όσο και στη πλευρά των ατομικών συσκευών των χρηστών. Κλείνοντας επιδόσεις μέχρι σήμερα όπως είδαμε ήταν σε κάποιο βαθμό ικανοποιητικές. Παρ' όλα αυτά οι απαιτήσεις είναι ακόμα πιο υψηλές και στην κατεύθυνση αυτή θα πρέπει να κινηθεί η μελλοντική εργασία, η οποία και περιέχεται στο επόμενο κεφάλαιο. Επίσης πιο αναλυτικά συμπεράσματα για κάθε ενότητα πειραμάτων υπάρχουν τόσο στις επί μέρους παραγράφους όσο και μετά από κάθε διάγραμμα αποτελεσμάτων.

Κεφάλαιο 6: *Ανακεφαλαίωση*

Συνεισφορά - Συμπεράσματα

Σε αυτήν την εργασία ασχοληθήκαμε με την υλοποίηση ενός συστήματος για τις ανάγκες της ψηφιακής τηλεόρασης. Αφορμή στάθηκε το ευρωπαϊκό πρόγραμμα up-TV η αρχιτεκτονική του οποίου είναι βασισμένη στο σχήμα μεταδεδομένων του TV-Anytime φόρουμ. Στόχος ήταν η κατασκευή ενός συστήματος αυτόματης κατασκευής και αναπροσαρμογής του προφίλ καθώς και η διεξαγωγή πειραμάτων για τον έλεγχο της λειτουργίας του συστήματος καθώς και της απόδοσης των μηχανισμών του. Αναλυτικότερα οι στόχοι που καλύφθηκαν ήταν η εξής:

- Αξιοποίηση των ενεργειών που καταγράφονται κατά την αλληλεπίδραση του χρήστη με το σύστημα για την συναγωγή προτιμήσεων για κάθε πρόγραμμα. Επίσης δόθηκε η δυνατότητα για αξιοποίηση ενεργειών και σε τμήματα προγραμμάτων. Η καταγραφή του ιστορικού των ενεργειών καθώς και η περιγραφή των προγραμμάτων και των τμημάτων τους είναι σύμφωνη με το TVA σχήμα.
- Χρησιμοποίηση ρητών και μη ρητών προτιμήσεων χρηστών πάνω σε προγράμματα για την κατασκευή του προφίλ τους. Με βάση την TVA περιγραφή για τις προτιμήσεις φιλτραρίσματος και αναζήτησης οι δομές προφίλ που υποστηρίχθηκαν ήταν οι ακόλουθες:
    - Επίπεδη δομή με έναν μόνο FASP κόμβο
    - Ιεραρχική δομή βασισμένη σε κάποιο χαρακτηριστικό
    - Ιεραρχική δομή βασισμένη σε κάποιο πρότυπο ιεραρχίας δομημένης με βάση συγκεκριμένα κριτήρια
- Μηχανισμοί για τον συνδυασμό δύο δομών προφίλ με στόχο την αναπροσαρμογή ενός παλιού με ένα νέο προφίλ για κάποιον χρήστη. Η λειτουργικότητα υποστηρίχθηκε και για τις τρεις παραπάνω δομές προφίλ.

- Έλεγχος της ορθής λειτουργίας και της απόδοσης των διάφορων μηχανισμών που αναπτύχθηκαν παραπάνω με διεξαγωγή πειραμάτων.
- Μελέτη συμπεριφοράς του συστήματος για τις διαφορετικές παραμέτρους των υποσυστημάτων.

Κύρια διαφορά του από υπάρχοντα συστήματα είναι η χρήση αναλυτικών μεταδεδομένων τόσο για την κατασκευή του προφίλ όσο και για τη συσχέτιση του προφίλ με τα μεταδεδομένα των προγραμμάτων. Το προφίλ κάθε χρήστη περιγράφεται αναλυτικά από το TVA με όρους που επιτρέπουν την άμεση συσχέτισή του με τα μεταδεδομένα των προγραμμάτων.

Με αυτόν το τρόπο αντί της χρήσης παραδοσιακών τεχνικών όπως συνεργατικό φιλτράρισμα(collaboration filtering) είναι δυνατή η απευθείας συσχέτιση του προφίλ με τα προγράμματα μέσω μοντέλων όπως το Extended Boolean. Πρόκειται για μεθοδολογίες με πολύ μικρότερες χρονικές απαιτήσεις. Παρ' όλα αυτά χαρακτηρίζονται από ένα σύνολο παραμέτρων η ρύθμιση των οποίων είναι κρίσιμη για την απόδοση του συστήματος. Στα πλαίσια της διπλωματικής έγινε μια πρώτη προσπάθεια για την μελέτη της επίδρασης τους και κυρίως μια προσέγγιση για την εκμετάλλευση τους για την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τη σημασία των διάφορων χαρακτηριστικών που εμφανίζονται στο προφίλ του χρήστη. Σε συστήματα που λαμβάνεται υπόψη η συνολική προτίμηση ενός χρήστη για κάθε πρόγραμμα υποκρύπτεται η υπόθεση πως η προτίμηση του χρήστη για το πρόγραμμα οφείλεται ουσιαστικά στα χαρακτηριστικά του προγράμματος από τα οποία το καθένα συμβάλει στο δικό του βαθμό. Από εκεί και πέρα όλες οι εκτιμήσεις γίνονται χρησιμοποιώντας αυτήν την προτίμηση που αντιπροσωπεύει το πρόγραμμα σαν σύνολο. Από την άλλη πλευρά στο σύστημα που αναπτύξαμε παρέχεται η δυνατότητα για αξιοποίηση της προτίμησης για μεμονωμένα χαρακτηριστικά προγραμμάτων. Οι δυνατότητες που ανοίγονται είναι μεγάλες καθώς πλέον είναι εφικτό να προσδιοριστούν συγκεκριμένες κατηγορίες χαρακτηριστικών προγραμμάτων που επιδρούν περισσότερο στην δημιουργία μιας συνολικής προτίμησης για κάποιο πρόγραμμα. Αντίστοιχα κατά τη προσπάθεια εκτίμησης της προτίμησης ενός χρήστη για νέα προγράμματα είναι δυνατό να χρησιμοποιηθούν μεμονωμένα χαρακτηριστικά ή ομάδες χαρακτηριστικών και όχι απλά ολόκληρα προγράμματα. Το μοντέλο του TVA προσεγγίζει περισσότερο τη συμπεριφορά Information Retrieval τεχνικών παρουσιάζοντας πιο διαζευκτική συμπεριφορά. Μετά και

από τα πειράματα που πραγματοποιήσαμε διαπιστώσαμε πως το θέμα της χρήσης διαζεύξεων ή συζεύξεων καθώς και το σε ποια σημεία του προφίλ χρησιμοποιείται αυτή χρειάζεται πολύ περισσότερη μελέτη. Με αυτόν τον τρόπο θα γίνει εφικτή και καλύτερη αποτύπωση συσχετίσεων μεταξύ συγκεκριμένων χαρακτηριστικών του προφίλ. Οι συσχετίσεις αυτές είναι που χαρακτηρίζουν κυρίως τα συστήματα εξόρυξης δεδομένων(Data Mining). Αυτό θα επέτρεπε στο σύστημα την εύκολη καταγραφή προτύπων προτιμήσεων από συνδυασμό χαρακτηριστικών, κάτι που μπορεί να εκφράσει με πιο ολοκληρωμένο τρόπο πρότυπα συμπεριφοράς από την πλευρά του χρήστη.

Συνοψίζοντας, μετά την ολοκλήρωση αυτής της εργασίας, υπάρχει ολοκληρωμένο ένα σύστημα κατασκευής και αναπροσαρμογής προφίλ χρηστών για περιβάλλον ψηφιακής τηλεόρασης. Επίσης έγινε μελέτη της λειτουργίας του συνολικού συστήματος. Οι παράμετροι λειτουργίας του είναι πολλές, ενώ η μορφή του προφίλ επιβάλλει τη χρήση νέων μεθοδολογιών σε έναν χώρο σαν αυτό της ψηφιακής τηλεόρασης. Η μελλοντική εργασία και επεκτάσεις είναι πολλές και κάποια πρώτα σημεία συνοψίζονται στην επόμενη παράγραφο.

Μελλοντική Εργασία

Στη συνέχεια θα παρουσιάσουμε κάποιες πρώτες σκέψεις σχετικά με τις ανάγκες για την ολοκλήρωση και βελτίωση του συστήματος όπως αυτό είναι σήμερα. Τα σημεία που θα αναφέρουμε είναι με τη σειρά παρουσίασης των υποσυστημάτων όπως ήταν αυτή στο υπόλοιπο κείμενο.

Ξεκινώντας από την χρήση των ενεργειών του χρήστη για την εκτίμηση των προτιμήσεων του, μια πιο αναλυτική προσέγγιση είναι απαραίτητη. Με βάση και την αντίστοιχη δουλειά που έχει γίνει για της ενέργειες των χρηστών σε εφαρμογές για το διαδίκτυο θα πρέπει να γίνει μελέτη της αλληλουχίας των ενεργειών και της σημασίας που μπορεί να έχουν για την προτίμηση του χρήστη. Επίσης πρότυπα αλληλουχίας ενεργειών(patterns) με συγκεκριμένη σημασία θα μπορούσαν να ληφθούν υπόψη κατά την εξέταση των ενεργειών. Για τις ενέργειες θα πρέπει να γίνει μια μελέτη των βαρών που τους αποδίδονται. Για όλα τα παραπάνω σημαντικό ρόλο θα έπαιζε η δυνατότητα να διεξαχθούν πειράματα με πραγματικά δεδομένα από καταγραφή ενεργειών χρηστών κατά τη χρήση του συστήματος.

Σχετικά με τα προφίλ των χρηστών χρειάζεται μελέτη τόσο θεωρητική όσο και πειραματική για την περίπτωση των ιεραρχικών προφίλ ώστε να βρεθεί ο καλύτερος τρόπος εκτίμησης αρχικά των προτιμήσεων για τα διάφορα χαρακτηριστικά του προφίλ και στη συνέχεια για την καλύτερη αλγεβρική μετάφραση του μοντέλου κατά τη συσχέτιση με τα μεταδεδομένα των προγραμμάτων.

Τέλος θα πρέπει να γίνουν πειράματα με δεδομένα που να αφορούν όχι μόνο ταινίες αλλά ένα πολύ μεγαλύτερο σύνολο από τις κατηγορίες προγραμμάτων που υποστηρίζει το TVA. Επίσης θα μπορούσε να γίνει συλλογή δεδομένων από πραγματικούς χρήστες με πληροφορία για την προτίμησή τους πάνω και σε μεμονωμένα χαρακτηριστικά των προγραμμάτων. Ο γενικότερος στόχος είναι ο όσο το δυνατόν καλύτερος συνδυασμός των χαρακτηριστικών που αντιπροσωπεύουν τις προτιμήσεις του χρήστη στο προφίλ του.

Βιβλιογραφία

[1] M. *Claypool, D. Brown, P. Le and M. Waseda.* Inferring User Interest*, IEEE Internet Computing.*

[2] *D. Oard and J. Kim*. Modeling information content using observable behavior, *Proceedings of ASIST 2001 Annual Meeting*, November 3-8, Washington D.C.

[3] *D. Nichols.* Implicit rating and filtering, *DELOS Workshop 1997*.

[4] *P. Chan.* A non-invasive learning approach to building web user profiles. KDD-99 Workshop on Web analysis and User Profiling.

[5] *G. Adomavicius and A. Tuzhilin.* Using Data Mining to build customer profiles, *IEEE Computer, Vol. 34*.

[6] *L. Ardissono, L. Console and I. Torre.* An adaptive system for the personalized access to news, *AI Communications, Vol 14*.

[7] *P. Buono, M. Constabile, S. Guida, A. Piccinno, G. Tesoro.* Integrating user data and collaborative filtering in a web recommendation system, *UM2001*.

[8] *A. L. Buczak, J. Zimmerman and K. Kurapati.* Personalization: Improving ease-of-use, trust and accuracy of a TV show recommender, *Proceedings of the AHB'2002 Workshop on Personalization in Future TV*.

[9] *L. Ardissono, F. Portis and P. Torasso.* Architecture of a system for the generation of personalized Electronic Program Guides, *UM2001 Workshop on Personalization in Future TV*.

[10] *B. Smyth and P. Cotter. Surfing the Digital Wave.* Generating Personalized TV Listings using Collaborative, Case-Based Recommendation, *ICCBR 1999*.

[11] *S. Niiranen, A. Tampere and S. Kalli.* Agent-Based Personalization in digital television. P. Baudisch and L. Brueckner. TV Scout: Lowering the entry barrier to personalized TV program recommendation, *Proceedings of the AHB'2002 Workshop on Personalization in Future TV*.

[12] *P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl.* GroupLens: An Open Architecture, for Collaborative Filtering of News, *Proceedings of the 1994 Computer Supported Collaborative Work Conference.*

[13*] K. Yu, Z. Wen, X. Xu, M. Ester*. Feature Weighting and Instance Selection for Collaborative Filtering, *An International Journal, Springer, 2002*.

[14] Text of ISO/IEC 15938-5 *Information Technology - Multimedia content description interface - Part 5 Multimedia Description Schemes*, 2001.

[15] The TV-Anytime Forum **,** *Specification Series: S-3 on: Metadata (Normative)*, www.tvanytime.org, June 2002.

[16] *Γ. Κοτόπουλος.* «Σύστημα διήθησης TVA περιεχομένου προγραμμάτων Ψηφιακής Τηλεόρασης σύμφωνα με τα ενδιαφέροντα των χρηστών» Διπλωματική, Πολυτεχνείο Κρήτης 2003.