ΣΥΣΤΗΜΑ ΣΥΣΤΑΣΕΩΝ ΓΙΑ ΔΙΑΔΡΑΣΤΙΚΗ ΕΞΕΡΕΥΝΗΣΗ ΣΧΕΣΙΑΚΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από την

ΚΩΛΕΤΣΟΥ ΕΥΤΥΧΙΑ

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΟ ΛΟΓΙΣΜΙΚΟ

Ιούνιος 2010

ΑΦΙΕΡΩΣΗ

Σε όλους εκείνους που στέκονται δίπλα μου
πάντα με ένα ζεστό χαμόγελο.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση αυτής της εργασίας αισθάνομαι την ανάγκη να ευχαριστήσω όλους εκείνους που με στήριξαν και συνέβαλαν με τον δικό τους τρόπο στην περάτωσή της.

Πρώτα από όλους, ευχαριστώ θερμά την επιβλέπουσα καθηγήτρια μου κα Ευαγγελία Πιτουρά, αρωγό σε αυτή μου την προσπάθεια, που με τον δικό της τρόπο κατάφερνε συνεχώς να με διδάσκει, να με προτρέπει, να με ενθαρρύνει.

Ακόμη, υπάρχει ένας μεγάλος αριθμός από ανθρώπους, φίλους και συνεργάτες, που μου στάθηκαν όλο αυτό το διάστημα και τους οφείλω ένα μεγάλο ευχαριστώ. Όλοι αυτοί, και ο καθένας ξεχωριστά, με υποστήριζαν και με εμψύχωναν, με πίστεψαν και με προέτρεπαν, μου έδειχναν το δρόμο να προχωρήσω.

Τέλος θα ήθελα να ευχαριστήσω τους δικούς μου ανθρώπους, αυτούς που στάθηκαν δίπλα μου, με κατανόηση και υπομονή, και με στήριξαν όλο αυτό το διάστημα.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Ευτυχία Κωλέτσου του Γεωργίου και της Αγγελικής.
MSc Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούνιος, 2010.
Σύστημα Συστάσεων για Διαδραστική Εξερεύνηση Σχεσιακών Βάσεων Δεδομένων
Επιβλέπουσα: Ευαγγελία Πιτουρά.

Τα συστήματα των σχεσιακών βάσεων δεδομένων είναι πολύ δημοφιλή λόγω του ότι παρέχουν τη δυνατότητα επεξεργασίας πολύπλοκων ερωτημάτων, με αποτέλεσμα να επιτρέπουν στο χρήστη ανάκτηση ενδιαφέρουσας πληροφορίας. Παρόλα αυτά, καθώς οι βάσεις δεδομένων, ολοένα και αυξάνονται, όσον αφορά τον όγκο τους, και γίνονται προσβάσιμες συνεχώς σε ένα όλο και πιο διαφοροποιημένο και λιγότερο τεχνικά-καταρτισμένο κοινό, μία νέα μορφή αλληλεπίδρασης προσανατολισμένη σε *συστάσεις* φαίνεται να είναι ελκυστική και χρήσιμη.

Οι *συστάσεις* απασχολούν πλέον σημαντικά ερευνητικά πεδία, δεδομένου ότι αφθονούν σε πρακτικές εφαρμογές που βοηθάνε τους χρήστες να διαχειρίζονται υπερβολικά μεγάλο όγκο δεδομένων και παρέχουν προσωποποιημένες συστάσεις, περιεχόμενα και υπηρεσίες σε αυτούς. Για παράδειγμα, οι συστάσεις μπορούν να βρουν εφαρμογή σε ηλεκτρονικά καταστήματα, όπως e-βιβλιοπωλεία κλπ, για να αξιολογούν ποια προϊόντα ή υπηρεσίες ενδιαφέρουν τους χρήστες με σκοπό στη συνέχεια να τους προτείνουν σχετικά προϊόντα ή υπηρεσίες που θα ήθελαν να αγοράσουν.

Παρακινούμενοι από τον τρόπο που τα συστήματα συστάσεων λειτουργούν, θεωρούμε στις σχεσιακές βάσεις δεδομένων, ως συστάσεις προς τους χρήστες ακόμη και πλειάδες που δεν ανήκουν στο αποτέλεσμα της ερώτησής τους, αλλά που πιθανότατα είναι ενδιαφέρουσες για αυτούς.

Ονομάζουμε αυτά τα επιπλέον ανακτώμενα αποτελέσματα '*You May Also Like*' ή YMAL αποτελέσματα. Τα YMAL αποτελέσματα είναι ιδιαίτερα χρήσιμα διότι επιτρέπουν στους χρήστες να ανακαλύψουν επιπλέον πλειάδες της βάσης δεδομένων που ενδεχομένως να αγνοούσαν.

Υπολογίζουμε τα YMAL αποτελέσματα βασισμένοι στην προσέγγιση της *τρέχουσας κατάστασης* (***current-state approach***), η οποία εκμεταλλεύεται το περιεχόμενο του αποτελέσματος του τρέχοντος ερωτήματος που τίθεται από το χρήστη, καθώς και το σχήμα της βάσης δεδομένων. Μελετάμε τέσσερις διαφορετικούς τρόπους για να υπολογίσουμε YMAL αποτελέσματα.

Σε αυτή τη διατριβή παρουσιάζουμε επιπλέον το YMAL σύστημα συστάσεων, που αναπτύχθηκε για την εφαρμογή της μεθόδου των YMAL αποτελεσμάτων, το οποίο προσφέρει στο χρήση αλληλεπίδραση μέσω ενός δυναμικού web περιβάλλοντος. Το YMAL σύστημα συστάσεων υλοποιήθηκε με χρήση της τεχνολογίας JSP σε συνδυασμό με την αντικειμενοστραφή γλώσσα προγραμματισμού Java, πάνω σε σχεσιακή βάση δεδομένων υλοποιημένη με MySQL. Τέλος, παρουσιάζουμε ενδεικτικά παραδείγματα ερωτήσεων και YMAL αποτελεσμάτων, καθώς και μία μελέτη απόδοσης των τεσσάρων διαφορετικών προσεγγίσεων που υλοποιήσαμε.

EXTENDED ABSTRACT IN ENGLISH

Koletsou G. Eftychia
MSc, Computer Science Department, University of Ioannina, Greece. June, 2010.
Recommendation System for Interactive Database Exploration
Thesis Supervisor: Pitoura Evaggelia.

The typical interaction of a user with a database system is by formulating queries. This interaction mode assumes that users are to some extent familiar with the content of the database and also have a clear understanding of their information needs. However, as databases get larger and accessible to a more diverse and less technically-oriented audience, a new "recommendation"-oriented form of interaction seems attractive and useful.

Motivated by the way recommenders work; we consider extending relational database systems with recommendation functionality. In particular, we propose that, along with the results of each query, the user gets additional recommended results of potential interest. We call such results "*You May Also Like*" or YMAL results for short. YMAL results are useful because they let users see other tuples in the database that they may be unaware of.

We focus on the *current-state approach* to computing YMAL results. This method explores both the database schema by expanding the original query through joins with appropriate other relations and the database content through value correlations.

We assume first, that there is no other information available other than a query *Q* posed by a user *u* and its result *R*(*Q*). Then, YMAL results can be computed based on either (i) *local analysis* of the intrinsic properties of the result *R*(*Q*) or (ii) *global analysis* of the properties of the database *D*. In both cases, we can exploit (i) the *content* and/or (ii) the *schema* of *R*(*Q*) or *D* respectively.

There are many directions for computing YMAL results along these axes. We discuss a suite of these different approaches. In order to compute the YMAL results for a posed query, we exploit either the content-based approach or the schema-based approach.

In the content-based approach, through the local-based analysis, after *R*(*Q*) has been computed, we examine the content of its tuples to locate common information patterns appearing in many of them. We then employ such information to retrieve and recommend tuples of the database that do not belong in *R*(*Q*) but exhibit similar behavior. Another option, through the global-based analysis, we base YMAL computation on properties of *D*, relying on the correlation of specific attribute values as well as their selectivity.

In the schema-based approach, through the local-based analysis, we expand the tuples of the result through dynamical joins between the subset that becomes from the result and other relations in *D*. Intuitively, in this way we add extra, possibly useful information to the result and search for common patterns in the expanded result tuples. On the other hand, through the global-based analysis, we expand the tuples of the result through joins that we have pre-compute in *D*. In this way, correlation among relations can be used to direct the expansion of tuples in *R*(*Q*) in this view of the problem.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Πρόσφατα, η μοντελοποίηση και η διαχείριση προτεινόμενης πληροφορίας έχουν προσελκύσει την ιδιαίτερη προσοχή σε τομείς όπως είναι οι Βάσεις Δεδομένων, η Διαχείριση Γνώσης και η Ανάκτηση Πληροφορίας. Αυτό το ενδιαφέρον πηγάζει από το γεγονός ότι ολοένα και περισσότεροι μη ειδικευμένοι χρήστες έρχονται σε επαφή με τεράστιες συλλογές δεδομένων, χωρίς, κατά κανόνα, να έχουν μια σαφή άποψη για το περιεχόμενο ή για τη δομή της πληροφορίας που αναζητούν. Συνεπώς, συχνά προσπαθούν να ανακαλύψουν πληροφορία που ενδεχομένως θα τους είναι χρήσιμη.

1.1. Στόχοι της Διατριβής

Παρακινούμενοι από τον τρόπο που λειτουργούν τα συστήματα συστάσεων, και προσαρμόζοντάς τα στις σχεσιακές βάσεις δεδομένων, στοχεύουμε στη δημιουργία συστάσεων προς τους χρήστες.

Οι *συστάσεις* (***recommendations***) βοηθούν στο να λαμβάνουν οι χρήστες περαιτέρω πληροφορία, που ενδεχομένως να τους είναι χρήσιμη, χωρίς να διαθέτουν απαραίτητα την κατάλληλη εμπειρία ή γνώση για να διατυπώσουν την κατάλληλη ερώτηση. Συνήθως, οι συστάσεις υπολογίζονται με βάση την προηγούμενη συμπεριφορά του ίδιου ή άλλων χρηστών του συστήματος. Σε αυτήν την περίπτωση οι συστάσεις μπορεί να αφορούν πλειάδες δημοφιλής σε χρήστες με παρόμοια συμπεριφορά ή πλειάδες που μοιάζουν με πλειάδες που προτίμησε ο ίδιος χρήστης στο παρελθόν.

Σε αυτήν την εργασία, ακολουθούμε μια διαφορετική προσέγγιση για τον υπολογισμό συστάδων. Η προσέγγιση μας βασίζεται στην επεξεργασία της ερώτησης του χρήστη και την παραγωγή επιπλέον γνώσης μέσα από το σχήμα και το περιεχόμενο της ίδιας της βάσης δεδομένων.

Συγκεκριμένα, οι συστάσεις αυτές θα αποτελούνται από επιπλέον ανακτώμενες *πλειάδες* (tuples), σε σχέση με εκείνες που αρχικά ανακτήθηκαν από τα ερωτήματα των χρηστών, οι οποίες παρουσιάζουν ενδιαφέρον. Για παράδειγμα, στην ερώτηση '*Σε ποιες ταινίες έχει παίξει ο* ***Johnny Depp***', θα μπορούσαμε να προτείνουμε επιπλέον ποιοι είναι οι δημοφιλέστεροι *ρόλοι* τους οποίους έχει παίξει ο ***Johnny Depp***, ή να πάμε ένα βήμα πιο μακριά προτείνοντας ταινίες που έχει σκηνοθετήσει ο ***Tim Burton***, αφού οι εγγραφές ***Johnny Depp*** και ***Tim Burton*** εμφανίζονται επαναληπτικά πολλές φορές συσχετισμένες με τις ίδιες *ταινίες* μέσα στη βάση δεδομένων.

Αποκαλούμε τα επιπλέον ανακτώμενα αποτελέσματα ως '***You May Also Like***' ή YMAL αποτελέσματα. Μέσω της ανάπτυξης ενός ολοκληρωμένου δυναμικού συστήματος συστάσεων, επιδιώκουμε να προσφέρουμε στους χρήστες YMAL αποτελέσματα επιτρέποντας τους να ανακαλύψουν επιπλέον πλειάδες της βάσης δεδομένων που ενδεχομένως να αγνοούσαν πιο πριν. Στόχος είναι τα YMAL αποτελέσματα να χαρακτηρίζονται από *διαφορετικότητα* και *κάλυψη*, όπως θα δούμε στη συνέχεια.

Προσεγγίζουμε αυτό το στόχο κινούμενοι γύρω από το γενικό άξονα της *τρέχουσας κατάστασης* (***current-state***), προσέγγιση η οποία εκμεταλλεύεται το περιεχόμενο του αποτελέσματος του τρέχοντος ερωτήματος που τίθεται από το χρήστη, καθώς και το σχήμα της βάσης δεδομένων. Βάσει αυτής της προσέγγισης, υλοποιούμε τέσσερις διαφορετικές περιπτώσεις ανάκτησης YMAL αποτελεσμάτων και συγκρίνουμε τα ανακτώμενα αποτελέσματα, προσπαθώντας να οδηγηθούμε σε χρήσιμα συμπεράσματα.

Η επέκταση των αποτελεσμάτων ερωτημάτων σε βάσης δεδομένων μέσω συστάσεων έχει επίσης προταθεί μέσα από δύο πρόσφατες δουλείες των [6] και [8]. Το [8] προτείνει ένα πλαίσιο και μία σχετιζόμενη με αυτό μηχανή για τη δηλωτική προδιαγραφή της διαδικασίας συστάσεων, ενώ οι συστάσεις στο [6] βασίζονται στην προηγούμενη συμπεριφορά παρόμοιων χρηστών. Στην εν λόγω εργασία, διευθετούμε μία συγκεκριμένη διαδικασία συστάσεων, η οποία προτείνει αποτελέσματα που σχετίζονται άμεσα με το ερώτημα που τίθεται από τον εκάστοτε χρήστη, και προτείνουμε μεθόδους για τη δημιουργία τέτοιων συστάσεων.

1.2. Δομή της Διατριβής

Στη συνέχεια αυτής της διατριβής, παρουσιάζουμε τις κατηγοριοποιήσεις και τον τρόπο λειτουργίας των συστημάτων συστάσεων, έτσι όπως περιγράφονται μέσα από σχετικές ερευνητικές εργασίες (Κεφάλαιο 2). Στη συνέχεια, παρουσιάζουμε την αρχιτεκτονική του συστήματος μας, καθώς και τις διαφορετικές προσεγγίσεις στις οποίες εστιάσαμε για την υλοποίηση του (Κεφάλαιο 3). Επιδεικνύουμε διάφορα πειραματικά αποτελέσματα που προκύπτουν μέσω της λειτουργίας του συστήματός μας, καταλήγοντας σε ενδιαφέροντα συμπεράσματα (Κεφάλαιο 4). Τέλος, ανακεφαλαιώνουμε, τονίζοντας τα σημαντικά στοιχεία που προέκυψαν, και θέτοντας νέους στόχους για μελλοντική εργασία (Κεφάλαιο 5).

ΚΕΦΑΛΑΙΟ 2. ΣΥΣΤΑΣΕΙΣ ΓΙΑ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Το συγκεκριμένο κεφάλαιο αφορά στην ανάδειξη των τεχνικών μεθοδολογιών και των μοντέλων, επί των οποίων βασίστηκε η παρούσα εργασία, καθώς και στην συνοπτική παρουσίαση άλλων επιστημονικών έργων που σχετίζονται με το εν λόγω αντικείμενο ή με παραπλήσιες θεματικές περιοχές.

2.1. Συστήματα Συστάσεων

Τα *συστήματα συστάσεων* (*recommender systems*) περιγράφουν συνολικά τις ευφυείς τεχνικές που παρέχουν εξατομικευμένες υπηρεσίες, γνώσεις ή συμβουλές, οι οποίες εκτιμάται ότι θα ικανοποιήσουν τις απαιτήσεις του χρήστη. Τα εν λόγω συστήματα δραστηριοποιούνται πολύ στο χώρο του ηλεκτρονικού επιχειρείν μια και οι επιτυχημένες συστάσεις ή προβλέψεις για το χρήστη θα ωφελήσουν αντίστοιχα και τις επιχειρήσεις που τις διαθέτουν. Ήδη πολλά ηλεκτρονικά καταστήματα (Amazon1, CdNow2) χρησιμοποιούν τεχνικές συστάσεων.

Τα νέα ηλεκτρονικά μέσα προσφέρουν και νέες ευκαιρίες στην ανάπτυξη των συστημάτων που προσαρμόζονται στα μεταβαλλόμενα συμφέροντα των χρηστών κατά τη διάρκεια του χρόνου. Τα συστήματα συστάσεων βασίζονται σε αλγορίθμους οι οποίοι λαμβάνουν ως είσοδο τα χαρακτηριστικά και τις προτιμήσεις των χρηστών, τις σχέσεις μεταξύ αυτών καθώς και τις πληροφορίες για τα διαθέσιμα αντικείμενα και στη συνέχεια αξιολογούν, ταξινομούν και επιλέγουν αντικείμενα με κριτήριο το εκτιμώμενο «ενδιαφέρον» ενός χρήστη για αυτά. Μέσα από την εργασία των [3]

γίνεται μία προσπάθεια ομαδοποίησης των διαφορετικών προσεγγίσεων για την επίτευξη του υπολογισμού αποτελεσμάτων που μπορούν να χρησιμοποιηθούν από συστήματα συστάσεων, και ενδεχομένως να είναι ενδιαφέροντα για τον χρήστη.

2.1.1. Ορισμοί

Οι κυρίαρχες οντότητες που εμφανίζονται σε ένα σύστημα συστάσεων είναι ο *χρήστης* ***(user)*** και το *αντικείμενο* ***(item)***. Με τον όρο χρήστης αναφερόμαστε σε κάθε άτομο που – παρέχοντας πιθανώς προσωπικές πληροφορίες, το ιστορικό του ενδιαφέροντός του για κάποια αντικείμενα ή τη γνώμη/βαθμονόμηση του για αυτά – αναμένει ως ανταπόδοση να λαμβάνει εξατομικευμένες συστάσεις για νέα αντικείμενα που πιθανώς τον ενδιαφέρουν.

Για την επίτευξη του στόχου του, το σύστημα συστάσεων χρησιμοποιεί ένα *μοντέλο αναπαράστασης*. Μέσω του μοντέλου αναπαράστασης, το σύστημα συστάσεων διατηρεί και αξιοποιεί τα στοιχεία για τον/τους χρήστες, τα αντικείμενα και τις συσχετίσεις τους. Με βάση την επεξεργασία αυτών των στοιχείων το σύστημα συστάσεων είναι σε θέση να εξάγει δεδομένα με μορφή σύστασης ***(recommendation)*** ή *πρόβλεψης* ***(prediction)***.

Μια σύσταση περιλαμβάνει μια ομάδα αντικειμένων που το σύστημα προβλέπει ότι αποτελεί τις κορυφαίες προτιμήσεις του χρήστη. Η πρόβλεψη εκφράζει την εκτίμηση του συστήματος για την προτίμηση του χρήστη σε ένα συγκεκριμένο αντικείμενο. Ορίζοντας, λοιπόν, τον ενεργό χρήστη $u \in U$ (όπου U το σύνολο των χρηστών του συστήματος) ως τον χρήστη που επιθυμεί να λάβει ανταπόδοση από ένα σύστημα συστάσεων, έχουμε τις εξής εξόδους:

Πρόβλεψη = Εκτιμώμενη άποψη του χρήστη u για συγκεκριμένο αντικείμενο i που ανήκει στο σύνολο των αντικειμένων I της βάσης δεδομένων D.

Σύσταση = Λίστα κορυφαίων αντικειμένων που εκτιμάται ότι ενδιαφέρουν περισσότερο τον χρήστη u.

Η απόδοση ενός συστήματος συστάσεων ουσιαστικά συνάδει με τη δυνατότητα να παρέχει σωστή πρόβλεψη για το ενδιαφέρον του χρήστη.

2.1.2. Κατηγορίες Συστημάτων Συστάσεων

- Συστάσεις βασισμένες στο περιεχόμενο

Στις συστάσεις που βασίζονται στο περιεχόμενο η χρησιμότητα $\boldsymbol{U}(u, i)$ ενός αντικειμένου i για κάποιον χρήστη u υπολογίζεται βάση τον βαθμό χρησιμότητας $\boldsymbol{U}(u, i_j)$ που εκχώρησε στο παρελθόν ο χρήστης u σε αντικείμενα i_j που είναι «παρόμοια» με το αντικείμενο i. Έτσι, αν για παράδειγμα σε ένα σύστημα ενοικίασης ταινιών, ο χρήστης που μας ενδιαφέρει έχει βαθμολογήσει υψηλά στο παρελθόν ορισμένες ταινίες, το σύστημα θα ελέγξει τις εν λόγω ταινίες και μελλοντικά θα προτείνει στο χρήστη μόνο ταινίες που έχουν υψηλή συνάφεια με τις υψηλά βαθμολογούμενες ταινίες του χρήστη.

Η συγκεκριμένη προσέγγιση έχει τις ρίζες της στα ερευνητικά πεδία τόσο της ανάκτησης πληροφορίας όσο και στο φιλτράρισμα πληροφορίας [2]. Λόγω της πολύ καλής απόδοσης της μεθόδου σε αντικείμενα σύστασης που περιέχουν πληροφορία κειμένου, χρησιμοποιήθηκε ευρύτατα σε προφίλ χρηστών ώστε να συλλέγει πληροφορίες σχετικά με τις προτιμήσεις τους. Έτσι, η βασική ιδέα είναι ο εντοπισμός, από τη μία, των προφίλ των αντικειμένων (***Content***(*i*)), αποδιδόμενα ως *λέξεις-κλειδιά* (***keywords***), και από την άλλη, ο εντοπισμός των προφίλ των χρηστών (***ContentBasedProfile(u)***), υπολογίζοντας τα βάρη για τις λέξεις-κλειδιά για κάθε χρήστη, καταλήγοντας τέλος στον εντοπισμό της *συνημιτονικής τους ομοιότητας*:

$$\boldsymbol{U}(u, i) = \text{score}(\text{ContentBasedProfile}(u), \text{Content}(i)) \qquad \text{Εξ. 2.1}$$

Τα δύο μοντέλα που ακολουθούνται στις μεθόδους σύστασης είναι οι ***Bayesian*** *ταξινομητές* και οι *μετρήσεις πιθανοτήτων*.

Οι περιορισμοί που προκύπτουν στα συστήματα με μεθόδους βασισμένες στο περιεχόμενο είναι 1) η *περιορισμένη ανάλυση περιεχομένου*, λόγω ανεπάρκειας σε σύνολο χαρακτηριστικών, 2) η *υπερβολική εξειδίκευση*, που μπορεί να συστήσει συνεχώς μόνο πολύ παρόμοια είδη, 3) *ο νέος χρήστης*, για τον οποίο δεν παρέχονται επαρκείς πληροφορίες στο σύστημα ώστε να δομήσει σωστά το προφίλ του.

- *Συνεργατικές συστάσεις βασισμένες στο χρήστη*

Στις συνεργατικές συστάσεις γίνεται μία προσπάθεια πρόβλεψης της χρησιμότητας των αντικειμένων για έναν συγκεκριμένο χρήστη βασιζόμενη στα αντικείμενα που προηγουμένως έχουν βαθμολογηθεί από *άλλους χρήστες*. Η χρησιμότητα $\boldsymbol{U}(u, i)$ ενός αντικειμένου i για κάποιον χρήστη u υπολογίζεται βάση τον βαθμό χρησιμότητας $\boldsymbol{U}(u_j, i)$ που εκχωρήθηκε στο αντικείμενο i από όλους αυτούς τους χρήστες $u_i \in U$ και οι οποίοι είναι 'παρόμοιοι' με τον χρήστη u. Έτσι, για παράδειγμα σε ένα σύστημα ενοικίασης ταινιών, με σκοπό τη σύσταση ταινιών στον χρήστης που μας ενδιαφέρει, το σύστημα συνεργατικών συστάσεων ψάχνει να βρει τους ομότιμους (***peers***) του εν λόγω χρήστη, και έπειτα του προτείνει μόνο τις ταινίες που προτιμήθηκαν από τους ομότιμους του συγκεκριμένου χρήστη.

Μελετώντας τους αλγορίθμους για τις συνεργατικές συστάσεις θα μπορούσαμε να τους ομαδοποιήσουμε σε δύο γενικές κατηγορίες: τους ***memory-based*** (ή ***heuristic based***) και τους ***model-based*** [2]. Οι ***memory-based*** αλγόριθμοι κατ' ουσία είναι ευρεστικοί που κάνουν προβλέψεις βαθμολογίας βασιζόμενοι σε ολόκληρη την συλλογή των αντικειμένων που προηγουμένως βαθμολογήθηκαν από τους χρήστες. Αυτό σημαίνει ότι η τιμή μιας άγνωστης βαθμολογίας $r_{u,i}$ για κάποιον χρήστη u και ένα αντικείμενο i συνήθως υπολογίζεται ως ένα σύνολο από βαθμολογίες κάποιων άλλων χρηστών για το ίδιο αντικείμενο i.

$$r_{u,i} = aggr_{u' \in \bar{U}} r_{u',i} \qquad \text{Εξ. 2.2}$$

όπου $\bar{U}$ δηλώνει το σύνολο των N χρηστών που είναι οι πιο όμοιοι στο χρήστη u και οι οποίοι έχουν βαθμολογήσει το αντικείμενο i.

Διάφορες προσεγγίσεις έχουν χρησιμοποιηθεί για τον υπολογισμό της ομοιότητας ***sim*** (***u, u'***) μεταξύ χρηστών στα συστήματα συνεργατικών συστάσεων. Παρόλα αυτά, οι δύο πιο δημοφιλείς προσεγγίσεις είναι η ***correlation*** και η ***cosine-based***. Θεωρώντας ότι I_{xy} είναι το σύνολο των αντικειμένων που προτείνεται ταυτόχρονα από τους χρήστες *x* και *y*, στην ***correlation*** προσέγγιση η ομοιότητα υπολογίζεται σύμφωνα με τον τύπο:

$$sim(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})^2 \sum_{i \in I_{xy}} (r_{y,i} - \bar{r_y})^2}} \qquad \text{Εξ. 2.3}$$

ενώ, στην ***cosine-based*** προσέγγιση οι δύο χρήστες *x* και *y* θεωρούνται ως δύο διανύσματα σε ένα *m*-διάστατο χώρο, όπου $m=|I_{xy}|$. Έτσι η ομοιότητα μεταξύ των δύο διανυσμάτων μπορεί να μετρηθεί με τον υπολογισμό του *συνημίτονου* της μεταξύ τους γωνίας:

$$sim(x,y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|^2 \times \|\vec{y}\|^2} = \frac{\sum_{i \in I_{xy}} r_{x,i} \cdot r_{y,i}}{\sqrt{\sum_{i \in I_{xy}} r_{x,i}} \cdot \sqrt{\sum_{i \in I_{xy}} r_{y,i}}} \qquad \text{Εξ. 2.4}$$

Τα δύο μοντέλα που ακολουθούνται στις συνεργατικές συστάσεις είναι τα *μοντέλα* ***Cluster*** και τα ***Bayesian*** *δίκτυα*. Οι περιορισμοί που προκύπτουν σε αυτά τα συστήματα είναι 1) *ο νέος χρήστης*, για τον οποίο δεν παρέχονται επαρκείς πληροφορίες στο σύστημα ώστε να δομήσει σωστά το προφίλ του (παρόμοιο πρόβλημα όπως και στα συστήματα συστάσεων που βασίζονται στο περιεχόμενο), 2) *το νέο αντικείμενο*, λόγω του ότι ελάχιστα νέα αντικείμενα έχουν βαθμολογηθεί, και 3) *αραίωση*, ελάχιστα ζεύγη χρηστών έχουν επαρκώς αλληλοβαθμολογούμενα αντικείμενα ώστε να δημιουργηθεί μία παρόμοιο ομάδα μεταξύ τους.

- *Υβριδικές μέθοδοι συστάσεων*

Διάφορα συστήματα συστάσεων χρησιμοποιούν την *υβριδική προσέγγιση* [7] συνδυάζοντας τις συστάσεις που βασίζονται στο περιεχόμενο και τις συνεργατικές συστάσεις. Οι διαφορετικοί τρόποι που μπορεί να γίνει αυτός ο συνδυασμός αναφέρεται και σε κάθε μία από τις ακόλουθες υποκατηγορίες [2]:

Συνδυασμός ξεχωριστών συστάσεων: Ένας τρόπος για να χτίσει κανείς ένα υβριδικό σύστημα είναι να εφαρμόσει ξεχωριστά τα δύο προαναφερόμενα συστήματα. Υπάρχουν δύο διαφορετικά σενάρια για αυτό: (α) συνδυάζοντας τις βαθμολογίες που προέρχονται από ατομικά συστήματα συστάσεων μέσα σε μία τελική σύσταση χρησιμοποιώντας είτε ένα γραμμικό συνδυασμό των αξιολογήσεων είτε ένα σύστημα ψηφοφορίας, (β) χρησιμοποιώντας μία ατομική συνιστώσα, ανά πάσα στιγμή επιλέγοντας τη χρήση της 'καλύτερης' βασιζόμενοι σε μετρικές 'ποιότητας' συστάσεων.

Προσθήκη χαρακτηριστικών από μεθόδους που βασίζονται στο περιεχόμενο μέσα στα συνεργατικά μοντέλα: Αρκετά υβριδικά συστήματα συστάσεων βασίζονται σε παραδοσιακές συνεργατικές τεχνικές αλλά επίσης διατηρούν προφίλ βασισμένο στο περιεχόμενο για κάθε χρήστη. Η χρήση αυτών των προφίλ βοηθάει στον υπολογισμό ομοιότητας μεταξύ χρηστών. Επιπλέον, οι χρήστες μπορούν να προτείνουν ένα αντικείμενο άμεσα, και όχι μόνο όταν αυτό το αντικείμενο είναι υψηλά βαθμολογούμενο από άλλους χρήστες με παρόμοια προφίλ.

Προσθήκη χαρακτηριστικών από μεθόδους που βασίζονται στα συνεργατικά μοντέλα μέσα στα μοντέλα που βασίζονται στο περιεχόμενο: Η πιο γνωστή προσέγγιση σε αυτή την υποκατηγορία είναι η χρήση μερικών τεχνικών μείωσης χαρακτήρων σε μία ομάδα από προφίλ βασισμένα στο περιεχόμενο.

Ανάπτυξη ενός μοναδικού ενιαίου μοντέλου συστάσεων: Είναι η προσέγγιση που ακολουθείται από πολλούς ερευνητές τα τελευταία χρόνια. Ορισμένοι έχουν προτείνει τη χρήση του υβριδικού μοντέλου σε ένα ενιαίο κανόνα ταξινόμησης, άλλοι μία ενοποιημένη πιθανοκρατική μέθοδο για συνδυασμό των συνεργατικών και των βασισμένων στο περιεχόμενο συστάσεων, και άλλοι Bayesian μικτής-επίπτωσης παλινδρόμησης μοντέλα για την εκτίμηση παραμέτρων και την δημιουργία προβλέψεων.

2.1.3. Δυνατότητες επέκτασης στα συστήματα σύστασης

Σύμφωνα με όλες τις περιγραφές των συστημάτων σύστασης που μέχρι τώρα περιγράφηκαν σε αυτή την εργασία, γίνεται κατανοητό ότι οι μέχρι τώρα μέθοδοι συστάσεων χαρακτηρίζονται παράλληλα από διάφορους περιορισμούς κατά την άμεση εφαρμογή τους. Στο [2] προτείνονται διάφορες δυνατότητες επέκτασης των συστημάτων σύστασης.

Αρχικά, η συνολική *κατανόηση χρηστών και αντικειμένων*, φαίνεται να απασχολεί αρκετά τους συγγραφείς του άρθρου, οι οποίοι και προτείνουν τη συλλογή δεδομένων από τα προφίλ των χρηστών, όπως λέξεις-κλειδιά, δημογραφικά στοιχεία των χρηστών, εξόρυξη δεδομένων, ακολουθίες, και υπογραφές, για μία πιο καλή κατανόηση και σκιαγράφηση τους. Επιπλέον, αναφέρονται σε πρότυπα πλοήγησης (***navigational patterns***) στο Web, με σκοπό μία καλύτερη πρόβλεψη για τα προτεινόμενα Web sites.

Μία δεύτερη ιδέα είναι οι *επεκτάσεις του βασικού μοντέλου των τεχνικών σύστασης* με χρήση πιο πολλών μαθηματικών θεωριών, όπως αυτή των ***radial basis functions***. Οι συγγραφείς υποστηρίζουν ότι ένα από τα πλεονεκτήματα της εν λόγω θεωρίας είναι ότι έχει μελετηθεί εκτενώς στην προσεγγιστική θεωρία και οι θεωρητικές της ιδιότητες και χρήσεις έχουν κατανοηθεί καλά σε πολλές πρακτικές εφαρμογές.

Συνεχίζοντας, εκθέτουν την ιδέα των *πολυδιάστατων συστάσεων*, όπου προτείνουν τη συλλογή επιπλέον πληροφορίας κειμένου, όπως για παράδειγμα χρόνος, τόπος κτλ., διευρύνοντας την ιδέα των παραδοσιακών δισδιάστατων μεθόδων σύστασης, τύπου *Χρήστης x Αντικείμενο*. Έτσι, πλέον, προτείνεται η έννοια της συνάρτησης *χρησιμότητας* (ή βαθμολόγησης) πάνω σε ένα πολυδιάστατο χώρο $D_1 \times D_2 \times \ldots \times D_n$, ως $\boldsymbol{U}: D_1 \times \ldots \times D_n \to R$. Ακόμη, παραθέτουν την έννοια της ***reduction-based*** σύστασης, μία προβολή των πολυδιάστατων χαρακτηριστικών πάνω στο *Χρήστης x Αντικείμενο*.

Έπεται η ιδέα της *αξιολόγησης πολλαπλών κριτηρίων*, η οποία μπορεί να υλοποιηθεί με τους ακόλουθους τρόπους: (i) με την εύρεση Pareto βέλτιστων λύσεων, (ii)

λαμβάνοντας ένα γραμμικό συνδυασμό των πολλαπλών κριτηρίων και της μείωσης του προβλήματος σε ένα ενιαίου-κριτηρίου πρόβλημα βελτιστοποίησης, (iii) βελτιστοποιώντας τα πιο σημαντικά κριτήρια και μετατρέποντας άλλα κριτήρια σε περιορισμούς, (iv) βελτιστοποιώντας διαδοχικά ένα κριτήριο κάποια στιγμή, μετατρέποντας μία βέλτιστη λύση σε περιορισμό, και επαναλαμβάνοντας τη διαδικασία και για άλλα κριτήρια.

Πολλά συστήματα συστάσεων είναι διεισδυτικά υπό την έννοια ότι απαιτούν ρητή ανατροφοδότηση από το χρήστη και συχνά σε σημαντικό βαθμό τη συμμετοχή των χρηστών. Σύμφωνα με τους [2], ένας τρόπος για να διερευνηθεί το πρόβλημα της διείσδυσης είναι να ζητείται από κάθε νέο χρήστη να καθορίζει έναν βέλτιστο αριθμό αξιολογήσεων του συστήματος. Με αυτό τον τρόπο μπορεί να επιτευχθεί μία ισορροπία μεταξύ της ζήτησης περισσότερων ερωτήσεων προς τον χρήστη (ακρίβεια) και της μη-διείσδυσης (φιλικότητα προς το χρήστη). Μερικές ευρεστικές μέθοδοι *μη-διείσδυσης* προτείνονται σε αυτό το σημείο, από τους συγγραφείς, ως επεκτάσεις των ήδη υπαρχόντων. Πιο συγκεκριμένα, σε μία προσπάθεια καθορισμού και μέτρησης του οφέλους $B(n)$ της παροχής n αρχικών αξιολογήσεων αυξανόμενης ακρίβειας προβλέψεων βασιζόμενες σε αυτές τις αξιολογήσεις, χρειάζεται για να προσδιοριστεί ένας βέλτιστος αριθμός από αρχικές αξιολογήσεις n που θα μεγιστοποιεί την έκφραση $B(n)-C \cdot n$. Η βέλτιστη τιμή του n επιτυγχάνεται όταν οριακά οφέλη είναι ίσα με οριακά κόστη, δηλαδή όταν $\Delta B(n) = C$. Άλλη μία ενδιαφέρουσα ιδέα έγκειται στην ανάπτυξη οριακών μοντέλων κόστους, που είναι πιο προχωρημένα από το σταθερό μοντέλο κόστους, και που μπορούν ενδεχομένως να περιλαμβάνουν μία ανάλυση κόστους/οφέλους χρήσης τόσο έμμεσης όσο και ρητής αξιολόγησης των συστημάτων σύστασης.

Οι περισσότερες μέθοδοι σύστασης είναι μη-ευέλικτες, με συνέπεια ο τελικός χρήστης να μη μπορεί να προσαρμόσει τις συστάσεις σύμφωνα με τις τρέχουσες προσωπικές του ανάγκες. Προτείνεται, λοιπόν, η έννοια της *ευελιξίας* των συστημάτων αυτών μέσω της ***Recommendation Query Language*** (***RQL***).

Τέλος, άλλες επεκτάσεις που αναφέρονται στη βιβλιογραφία [2] έχουν να κάνουν με θέματα επεξηγηματικότητας, αξιοπιστίας, και επεκτασιμότητας που αφορούν τα συστήματα συστάσεων.

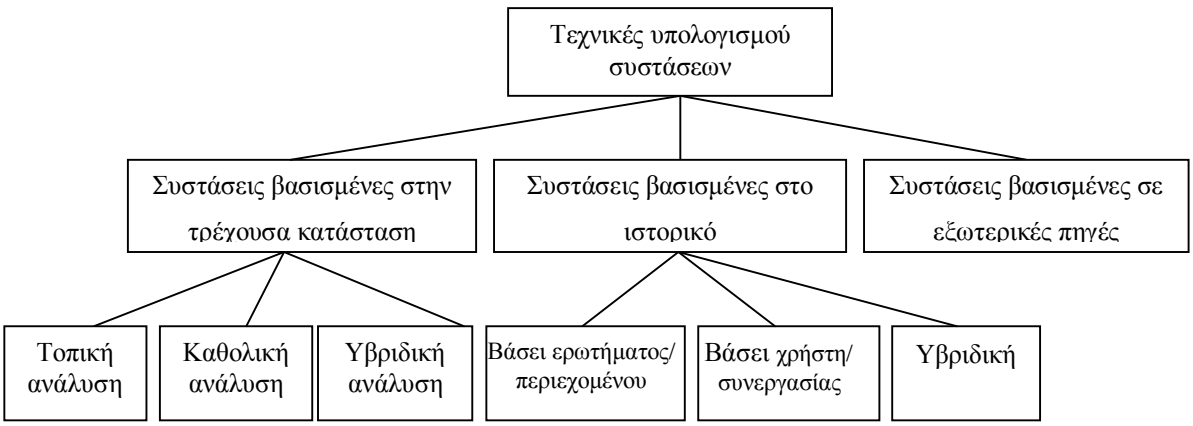
2.2. Συστήματα Συστάσεων για Σχεσιακές Βάσεις Δεδομένων

Μέσα από την εργασία [3] γίνεται μία προσπάθεια ομαδοποίησης των διαφορετικών προσεγγίσεων για την επίτευξη του υπολογισμού αποτελεσμάτων που μπορούν να χρησιμοποιηθούν από συστήματα συστάσεων, και ενδεχομένως να είναι ενδιαφέροντα για τον χρήστη. Αυτές οι προσεγγίσεις μπορούν να κατηγοριοποιηθούν σε τρεις κύριες κατηγορίες (Σχήμα 2.1):

Συστάσεις βασισμένες στην τρέχουσα κατάσταση (***Current-state***), που εκμεταλλεύονται το περιεχόμενο του αποτελέσματος του τρέχοντος ερωτήματος και το σχήμα της βάσης δεδομένων.

Συστάσεις βασισμένες στο ιστορικό (***History-based***), που εκμεταλλεύονται το ιστορικό των προηγούμενων τεθέντων ερωτημάτων στο σύστημα της βάσης δεδομένων, για παράδειγμα με τη χρήση καταγεγραμμένων ερωτήσεων.

Συστάσεις βασισμένες σε εξωτερικές πηγές (***External sources***), που εκμεταλλεύονται εξωτερικές πηγές πέραν της βάσης δεδομένων, όπως σχετικά δημοσιευμένα αποτελέσματα και αναφορές, σχετικές ιστοσελίδες, θησαυρούς ή οντολογίες.



Σχήμα 2.1 Ταξινόμηση Τεχνικών Υπολογισμού Συστάσεων

2.2.1. Συστάσεις βασισμένες στην τρέχουσα κατάσταση (Current-state)

Η αρχική υπόθεση των [3] σε αυτή την προσέγγιση είναι ότι δεν υπάρχει καμία άλλη πληροφορία διαθέσιμη εκτός από το ερώτημα *Q* που τέθηκε από το χρήστη *u* και το σύνολο των αποτελεσμάτων του ερωτήματος *R*(*Q*). Έπειτα, τα ενδεχομένως ενδιαφέροντα για τον χρήστη αποτελέσματα θα μπορούσαν να υπολογιστούν με βάση είτε (i) της *τοπικής ανάλυσης* των εγγενών ιδιοτήτων των αποτελεσμάτων του ερωτήματος *R*(*Q*), είτε (ii) της *καθολικής ανάλυσης* των ιδιοτήτων της βάσης δεδομένων *D*. Και στις δύο περιπτώσεις, μπορεί να γίνει εκμετάλλευση (i) του *περιεχομένου* και/ή (ii) του *σχήματος* του *R*(*Q*) ή της *D* αντίστοιχα. Η εργασία μας βασίζεται σε αυτή την προσέγγιση συστάσεων, την οποία και θα αναλύσουμε διεξοδικά στα επόμενα κεφάλαια.

2.2.2. Συστάσεις βασισμένες στο ιστορικό (History-based)

Οι συστάσεις που βασίζονται στο ιστορικό θεωρούν δεδομένο ότι υπάρχει πληροφορία από προηγούμενες διαδράσεις των χρηστών με τη βάση δεδομένων *D*. Οι εργασίες των [1], [2], [7] και [8] παρέχουν αρκετά πλήρεις συγκριτικές μελέτες των διαφορετικών προσεγγίσεων που υλοποιούνται με αυτόν τον τρόπο, οι οποίες σε γενικές γραμμές μπορούν να κατηγοριοποιηθούν ως εξής: Οι συστάσεις που βασίζονται στο περιεχόμενο (***content-based***) συστήνουν τα στοιχεία στο χρήστη βασισμένες στις περιγραφές των προηγουμένως αξιολογημένων στοιχείων. Οι [3] θεωρούν εναλλακτικά την προσέγγιση αυτή ως συστάσεις βασισμένες στο ερώτημα (***query-based***). Οι συνεργατικές συστάσεις (***collaborative***) συνδυάζουν τους χρήστες με πρόσωπα με παρόμοια ενδιαφέροντα και προβαίνουν στις συστάσεις βασισμένες σε αυτόν τον συνδυασμό. Στην εργασία των [3] η προσέγγιση αυτή αναφέρεται ως συστάσεις βασισμένες στο χρήστη (***user-based***). Οι υβριδικές προσεγγίσεις (***hybrid approaches***) εκμεταλλεύονται και τις δύο μεθόδους προκειμένου να εξαλείψουν τα μειονεκτήματα. Υπάρχουν διάφορες υβριδικές προσεγγίσεις. Είναι δυνατό να εφαρμοστούν οι δύο μέθοδοι χωριστά και να συνδυαστούν τα αποτελέσματα ή να ενσωματωθούν μερικά χαρακτηριστικά μιας μεθόδου στην άλλη. Ο πιο αποτελεσματικός τρόπος, εν τούτοις, φαίνεται να είναι η κατασκευή ενός ενοποιημένου προτύπου.

2.2.3. Συστάσεις βασισμένες σε εξωτερικές πηγές

Μέχρι τώρα έχουμε παρουσιάσει πως είναι δυνατόν να εντοπίσουμε και να συστήσουμε ενδιαφέροντα αποτελέσματα στο χρήστη εξερευνώντας εσωτερική πληροφορία της βάσης δεδομένων *D*. Για παράδειγμα, χρησιμοποιώντας μεταξύ τιμών των χαρακτηριστικών της ή και των ίδιων των σχέσεων που εμφανίζονται μέσα στη *D*, ακόμη και βασιζόμενοι σε προγενέστερη πληροφορία που έχουμε λάβει στο παρελθόν από παρόμοιους χρήστες και ερωτήματα. Παρόλα αυτά, υπάρχουν περιπτώσεις όπου οι σχέσεις μεταξύ δεδομένων των αντικειμένων δεν γίνονται άμεσα αντιληπτές στη *D*, ακόμη και αν είναι παρούσες. Στις μέρες μας, υπάρχει μια πληθώρα από χρήσιμη και καλά οργανωμένη πληροφορία στο Web, σε μορφή άρθρων, αναφορών και σχολίων που διατηρούνται σε συλλογικές αποθήκες γνώσεων όπως η Wikipedia [15] και το LibraryThing [16]. Σύμφωνα με τους [3] ανακτημένη πληροφορία από τέτοιου είδους εξωτερικές πηγές μπορεί επίσης να χρησιμοποιηθεί για τον υπολογισμό συστάσεων.

2.3. Προηγούμενη Δουλειά

Η ανάκτηση πληροφορίας με χρήση λέξεων-κλειδιών (***keywords***) είναι μία πολύ δημοφιλής μέθοδος, γιατί απαλλάσσει το χρήστη από επιπλέον γνώση είτε μίας γλώσσας ερωτήσεων είτε του σχήματος της βάσης δεδομένων για να βρει την πληροφορία που τον ενδιαφέρει. Το [4] είναι μία επιτυχημένη προσπάθεια προσαρμογής της ανάκτησης πληροφορίας με χρήση λέξεων-κλειδιών στις σχεσιακές βάσεις δεδομένων, διευκολύνοντας τον χρήστη στην αναζήτηση του και χωρίς να απαιτείται από αυτόν η γνώση των σχέσεων ή των χαρακτηριστικών που ενυπάρχουν μέσα στη βάση δεδομένων.

Έτσι, παρόμοια με τις μηχανές αναζήτησης, ο χρήστης δίνει μια σειρά από λέξεις-κλειδιά (***keyword search***) και παίρνει ως αποτέλεσμα τα σχετικά δεδομένα που υπάρχουν στη βάση, για παράδειγμα, τις πλειάδες (***tuples***) που περιέχουν τις λέξεις που αναζητεί. Επιπλέον, ένα άλλο σημαντικό χαρακτηριστικό, που προκύπτει μέσα από τις μεθόδους που ακολουθούνται στο [4], είναι το γεγονός ότι μπορεί κανείς να πάρει και πλειάδες που δεν ανήκουν απαραίτητα στην ίδια σχέση (***table***) αλλά και σε διαφορετικές σχέσεις συνδεόμενες με κάποιο ξένο κλειδί (***foreign key***).

Οι [5] προχωράνε ένα βήμα πιο πέρα αυτή τη δουλεία και προτείνουν μία *εξατομικευμένη αναζήτηση* με χρήση λέξεων-κλειδιών σε σχεσιακές βάσεις δεδομένων, χρησιμοποιώντας τις προτιμήσεις του χρήστη, ενώ τα αποτελέσματα των ερωτήσεων κλειδιών διατάσσονται με βάση το βαθμό προτίμησης τους από το χρήστη, τη σχετικότητά τους με το ερώτημα και δύο νέες μετρικές ποιότητας που αξιολογούν την ωφελιμότητα τους ως ένα σύνολο, γνωστό ως κάλυψη (***coverage***) και ποικιλομορφία (***diversity***). Οι συγγραφείς του [5], παρουσιάζουν ένα αλγόριθμο για δημιουργία ερωτημάτων προτιμήσεων που χρησιμοποιεί την εμφάνιση της λέξης-κλειδί στις προτιμήσεις για να κατευθύνει την συνένωση των σχετικών πλειάδων από τις πολλαπλές σχέσεις, υποδεικνύοντας παράλληλα έναν τρόπο μείωσης της πολυπλοκότητας αυτού του αλγορίθμου μέσω μοιραζόμενων υπολογιστικών βημάτων.

Λόγω του μεγάλου και συνεχώς αυξανόμενου όγκου δεδομένων που είναι πλέον διαθέσιμος για κάθε χρήστη, οι χρήστες συχνά δυσκολεύονται να βρουν την πληροφορία που αναζητούν εύκολα και γρήγορα, και ενδεχομένως να παραβλέπουν εν αγνοία τους ερωτήματα από τα οποία θα ανακτούσαν σημαντική πληροφορία. Ακόμη, διαφορετικοί χρήστες είναι πιθανό να έχουν διαφορετικά ενδιαφέροντα, ανάγκες και χαρακτηριστικά. Με στόχο τη βοήθεια των χρηστών σε αυτό το πλαίσιο, και επηρεασμένοι από τα συστήματα εξατομίκευσης, οι [6] προτείνουν επιπλέον πληροφορία στο χρήστη που ενδεχομένως να τον ενδιαφέρει. Η βασική ιδέα τους στηρίζεται στην παρακολούθηση των ερωτημάτων του χρήστη, την αναγνώριση των τμημάτων της βάσης που μπορεί να είναι ενδιαφέροντα για τη συγκεκριμένη ανάλυση δεδομένων, και την πρόταση ερωτημάτων που θα ανακτούν σχετικά δεδομένα. Μέσα από το [6] παρουσιάζουν μια αρχική πειραματική μελέτη βασισμένη σε ερωτήματα πραγματικών χρηστών που δείχνει ότι το πλαίσιο που προτείνουν (***QueRIE framework***) μπορεί να οδηγήσει τελικά χρήσιμα αποτελέσματα.

Η πληροφορία μέσα από το πλαίσιο ***QueRIE*** απορρέει ως εξής: τα ερωτήματα του ενεργού χρήστη διαβιβάζονται τόσο στο DBMS όσο και στην Μηχανή Συστάσεων (***Recommendation Engine***). Το DBMS επεξεργάζεται κάθε ερώτημα και επιστρέφει ένα σύνολο αποτελεσμάτων. Παράλληλα το ερώτημα αποθηκεύεται σε ένα ***Query***

Log. Η Μηχανή Συστάσεων συνδυάζει τις εισροές του τρέχοντος χρήστη με πληροφορία που έχει ήδη συγκεντρωθεί από τη βάση δεδομένων από αλληλεπιδράσεις προηγούμενων χρηστών, όπως καταγράφονται στο *Query Log*, και δημιουργεί ένα σύνολο από προτάσεις ερωτημάτων που επιστρέφονται στο χρήστη.

Κάτι αντίστοιχο με τους [6] προτείνεται και στην πρόσφατη εργασία των [8] όπου παρουσιάζεται το *FlexRecs* πλαίσιο. Το *FlexRecs* αποσυνδέει την περιγραφή μίας μεθόδου από την εκτέλεσή της και υποστηρίζει ευέλικτες συστάσεις πάνω από δομημένα δεδομένα. Μία μέθοδος συστάσεων περιγράφεται δηλωτικά ως μία παραμετροποιήσιμη ροή υψηλού επιπέδου που αποτελείται από κλασσικούς σχεσιακούς και νέους τελεστές για τη δημιουργία και σύνθεση συστάσεων. Επιπλέον, περιγράφονται τόσο μία μηχανή ευέλικτων συστάσεων που υλοποιεί το παρόν πλαίσιο όσο και παραδείγματα ροών καθώς και πειραματικά δεδομένα που δείχνουν τις δυνατότητες για σύλληψη και υλοποίηση πολλαπλών, παλιών ή καινοτόμων, μεθόδων συστάσεων.

Στην εργασία τους οι [9] εξετάζουν το πρόβλημα της κατάταξης των απαντήσεων ενός ερωτήματος στην βάση δεδομένων όταν επιστρέφονται πάρα πολλές πλειάδες. Πιο συγκεκριμένα, παρουσιάζουν μεθοδολογίες για να αντιμετωπίσουν το πρόβλημα των συνδετικών και σειριακών ερωτημάτων, προσαρμόζοντας και εφαρμόζοντας αρχές πιθανοκρατικών μοντέλων από την περιοχή της ανάκτησης πληροφορίας για δομημένα δεδομένα. Η λύση που προτείνουν οι συγγραφείς είναι ανεξάρτητη του πεδίου ορισμού και βασίζεται σε ανάλυση των δεδομένων, των στατιστικών και των συσχετίσεων. Αυτή η λύση είναι παρόμοια με τη βασισμένη στο ιστορικό και τη βασισμένη στο περιεχόμενο προσέγγιση. Η κύρια διαφοροποίηση από τη δική μας προσέγγιση, είναι ότι εμείς δε θεωρούμε ως πλειάδες προς σύσταση στο χρήστη εκείνες που ήδη βρίσκονται στο σύνολο του αποτελέσματος της αρχικής του ερώτησης, ενώ αντίθετα οι [9] θεωρούν πως πρέπει να γίνει κατάταξη των πλειάδων που βρίσκονται στο αρχικό σύνολο αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 3. YMAL-ΣΥΣΤΗΜΑ ΣΥΣΤΑΣΕΩΝ ΣΧΕΣΙΑΚΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

Σε αυτό το κεφάλαιο γίνεται περιγραφή τόσο του μοντέλου YMAL όσο και του συστήματος που υλοποιήθηκε. Βασιζόμαστε πάνω στην προσέγγιση της ανάκτησης και σύστασης πληροφορίας που προέρχεται από την τρέχουσα κατάσταση, και παρουσιάζουμε μία νέα μέθοδο για υπολογισμό.

3.1. Συστήματα Συστάσεων Βασισμένα στην Τρέχουσα Κατάσταση

Θεωρούμε μία βάση δεδομένων *D* και ένα σύνολο από χρήστες *U* που αλληλεπιδρούν μαζί της θέτοντας SPJ (*select-project-join*) ερωτήματα. Δοθέντος ενός ερωτήματος *Q*, μία τυπική βάση δεδομένων θα επέστρεφε ένα σύνολο από αποτελέσματα *R*(*Q*) με τη μορφή πλειάδων, πιθανώς παραγόμενων μέσω συνενώσεων διαφόρων σχέσεων στην *D*. Εκτός από το *R*(*Q*), θα θέλαμε να εντοπίσουμε και να προτείνουμε στους χρήστες ένα σύνολο από πλειάδες που θα τους ήταν εξίσου ενδιαφέρουσες. Καλούμε αυτό το σύνολο πλειάδων *'You May Also Like'* πλειάδες ή YMAL *αποτελέσματα* για συντομία. Συμβολίζουμε αυτό το σύνολο ως YMAL (*Q*).

Για τον υπολογισμό των YMAL αποτελεσμάτων υπάρχουν πολλές κατευθύνσεις. Στην τοπική ανάλυση, μετά τον υπολογισμό του *R*(*Q*), εξετάζεται το περιεχόμενο των πλειάδων του για τον εντοπισμό συχνά εμφανιζόμενων προτύπων πληροφορίας.

Έπειτα, γίνεται χρήση αυτής της πληροφορίας για την ανάκτηση και σύσταση από τη βάση δεδομένων πλειάδων που *δεν* ανήκουν στο $R(Q)$ αλλά παρουσιάζουν παρόμοια συμπεριφορά.

Από την άλλη πλευρά, στην προσέγγιση που βασίζεται στο σχήμα της βάσης, η ιδέα είναι να διευρυνθούν οι πλειάδες του αποτελέσματος *συνενώσεων* (*joins*) με άλλες σχέσεις (πίνακες) που παρουσιάζονται στο σχήμα της βάσης. Πιο αναλυτικά, σε αυτή την περίπτωση, προστίθεται επιπλέον πιθανώς χρήσιμη πληροφορία στο αποτέλεσμα και γίνεται αναζήτηση για συχνών προτύπων στις πλειάδες με το διευρυμένο πλέον αποτέλεσμα.

Στην περίπτωση της καθολικής ανάλυσης, ο υπολογισμός των ενδεχομένως ενδιαφερόντων για τον χρήστη YMAL αποτελεσμάτων βασίζεται στις ιδιότητες της D. Η προσέγγιση του περιεχομένου βασίζεται στις συσχετίσεις συγκεκριμένων τιμών των χαρακτηριστικών καθώς και στον βαθμό επιλεκτικότητας τους. Στην προσέγγιση που βασίζεται στο σχήμα της βάσης, για την άμεση επέκταση των πλειάδων στο $R(Q)$ μπορεί να χρησιμοποιηθεί η συσχέτιση μεταξύ των σχέσεων.

Οι υβριδικές μέθοδοι μπορούν ομοίως να υπολογιστούν με τον συνδυασμό της τοπικής και καθολικής ανάλυσης ή την πληροφορία που εκρέει από τις προσεγκίσεις βασισμένες στο περιεχόμενο και στο σχήμα κατά τη διάρκεια επεξεργασίας του αποτελέσματος του αρχικού ερωτήματος. Ο Πίνακας 3.1 δείχνει την ταξινόμηση των συστάσεων που βασίζονται στην τρέχουσα κατάσταση.

Πίνακας 3.1 Ταξινόμηση Συστάσεων Βασισμένων Στην Τρέχουσα Κατάσταση

Προσέγγιση \ Ανάλυση	Τοπική	Καθολική	Υβριδική
Βασισμένη στο περιεχόμενο	Οι πιο συχνές τιμές στο $R(Q)$	Οι πιο συχνές τιμές στη D	Συνδυασμός συχνών τιμών σε $R(Q)$ και D
Βασισμένη στο σχήμα	Απευθείας ενώσεις με βάσει τις συχνές τιμές του $R(Q)$	Απευθείας ενώσεις με βάσει το βαθμό συσχέτισης μεταξύ των σχέσεων στη D	Απευθείας ενώσεις με βάσει τις συχνές τιμές του $R(Q)$ και το βαθμό συσχέτισης μεταξύ των σχέσεων στη D

3.1.1. Τυπικός Ορισμός YMAL Συστάσεων

Γενικεύοντας, για να υπολογίσουμε τις YMAL συστάσεις, εξερευνούμε το περιεχόμενο και το σχήμα του εκάστοτε αποτελέσματος του ερωτήματος και της βάσης δεδομένων. Έτσι, δοθέντος ενός SPJ ερωτήματος Q της μορφής:

```
select A from R where P1 and P2
```

το `R` συμβολίζει ένα σύνολο σχέσεων στη D, το `A` ένα σύνολο χαρακτηριστικών $\{A_1,\ldots,A_n\}$ που ανήκουν στις σχέσεις της D, το `P1` τον συνδυασμό των συνθηκών συνένωσης για το Q και το `P2` τον συνδυασμό των συνθηκών επιλογής για το Q. Για την επεξεργασία των αιτημάτων σύστασης, προτείνουμε την επανεγγραφή του υποβαλλόμενου ερωτήματος Q μέσα από ένα σύνολο ερωτημάτων αναφερόμενα ως YMAL *ερωτήματα*. Ένα YMAL ερώτημα είναι της μορφής:

```
select B from S where P
```

3.1.2. Παρουσίαση Συστάσεων

Δεδομένου ότι δεν επιθυμούμε να παρουσιάσουμε στο χρήστη μία πληθώρα συστάσεων που ανακτώνται από το σύνολο των YMAL ερωτημάτων, περιορίζουμε αυτά τα αποτελέσματα σε L αποτελέσματα, όπου L ένα κατώφλι που ορίζεται από το σύστημα, έχοντας ως βασικό στόχο την παρουσίαση των πιο δημοφιλών ανάμεσα σε εκείνα που υπάρχουν στη D. Για τον λόγο αυτό, μπορούμε να επεκτείνουμε τη μορφή των YMAL ερωτημάτων ως εξής:

```
select B from S where P group by B order by count(*) desc limit L
```

Την αντίστοιχη μορφή χρησιμοποιήσαμε και στο Κεφάλαιο 4 όπου μετράμε τις επιδόσεις εφαρμογής του μοντέλου μας, θέτοντας το L ίσο με 10.

Στη συνέχεια θα παρουσιάζουμε τον μηχανισμό επανεγγραφής του ερωτήματος για τον υπολογισμό των `B`, `S` και `P`. Αυτός ο μηχανισμός βασίζεται είτε (i) στην τοπική ανάλυση των εγγενών ιδιοτήτων του αποτελέσματος $R(Q)$, είτε (ii) στην καθολική

ανάλυση των ιδιοτήτων της *D*. Τα αποτελέσματα των YMAL ερωτημάτων αποτελούν τις YMAL συστάσεις, δηλαδή το YMAL (*Q*).

3.2. Αρχιτεκτονική YMAL Συστήματος Συστάσεων

Τα συστήματα συστάσεων που βασίζονται στην ανάκτησης και σύστασης πληροφορίας από την τρέχουσα κατάσταση επεξεργάζονται το σύνολο της πληροφορίας που απορρέει από την αρχική ερώτηση *Q*, ώστε να καθορίσουν τα YMAL αποτελέσματα για τον χρήστη. Όπως αναφέρθηκε και πιο πάνω, σε αυτή την εργασία ακολουθήσαμε τόσο την τοπική όσο και την καθολική ανάλυση, για τον υπολογισμό των YMAL αποτελεσμάτων, προσεγγίζοντας το πρόβλημα και από τις δύο πλευρές.

Στο σημείο αυτό, θα αναφέρουμε τα κύρια στοιχεία της αρχιτεκτονικής του συστήματος μας. Μία πιο γενική απεικόνισή του, παρουσιάζεται στο Σχήμα 3.1, ενώ μία πιο αναλυτική περιγραφή του συστήματος θα παρουσιαστεί στις ακόλουθες ενότητες, με βάση τον διαχωρισμό για την εκάστοτε προσέγγιση.

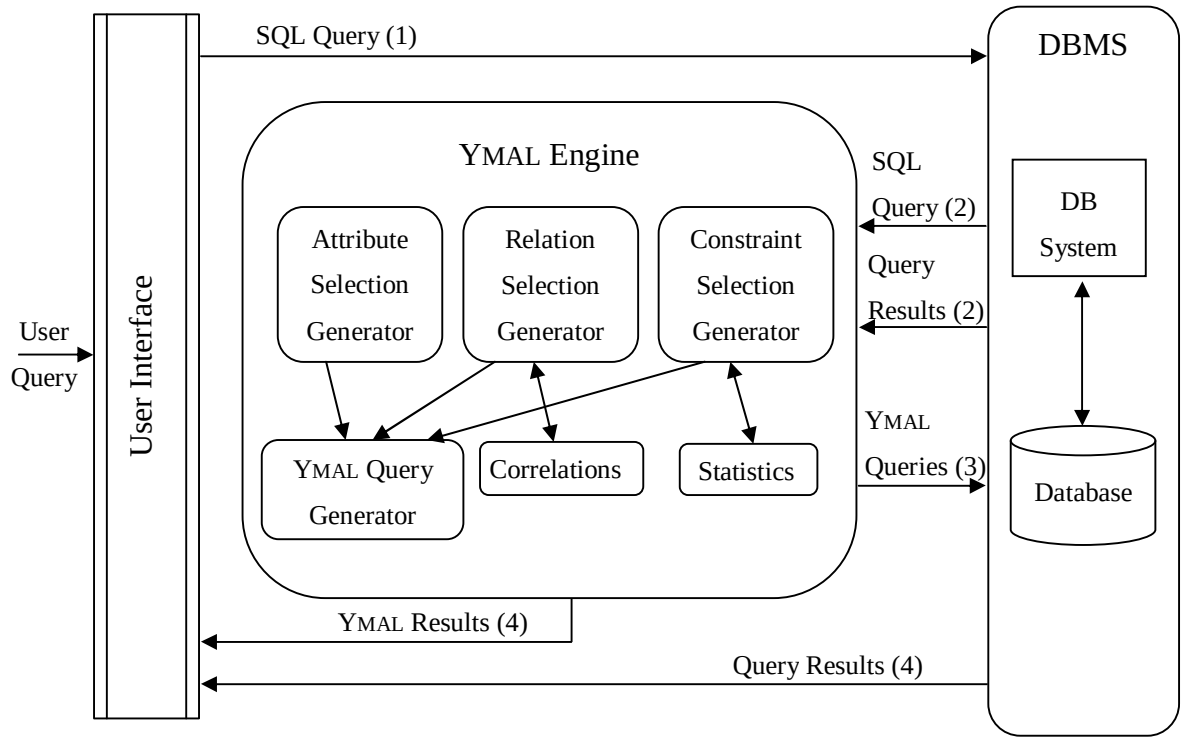
Καθώς υποβάλλεται ένα ερώτημα *Q*, υπολογίζουμε αρχικά το πραγματικό σύνολο των αποτελεσμάτων του. Χρησιμοποιώντας τα αποτελέσματα του ερωτήματος, δημιουργούμε ένα σύνολο από YMAL ερωτήματα επανεγγράφοντας το αρχικό ερώτημα *Q* που δόθηκε από το χρήστη. Τα αποτελέσματα αυτών των ερωτήσεων αντιστοιχούν στα YMAL αποτελέσματα.

Επιλογή Χαρακτηριστικών (***Attribute Selection Generator***): αυτή η συνιστώσα δέχεται ως είσοδο το υποβαλλόμενο ερώτημα *Q* και επιστρέφει τα χαρακτηριστικά του `B`, τα οποία θα εμφανιστούν ως γνωρίσματα του `select` των YMAL ερωτημάτων.

Επιλογή Σχέσεων (***Relation Selection Generator***): αυτή η συνιστώσα δέχεται ως είσοδο το υποβαλλόμενο ερώτημα *Q* και επιστρέφει τις σχέσεις `S` που θα εμφανιστούν στα γνωρίσματα του `from` των YMAL ερωτημάτων.

Επιλογή Περιορισμών (*Constraint Selection Generator*): αυτή η συνιστώσα δημιουργεί τις συνθήκες επιλογής που περιέχουν τα YMAL ερωτήματα, και δέχεται ως είσοδο τα *Q* και *R*(*Q*).

Δημιουργία YMAL ερωτήματος (*YMAL Query Generator*): Σε αυτό το βήμα, δημιουργούμε τα YMAL ερωτήματα. Για να εκτελέσουμε αυτή τη λειτουργία, χρησιμοποιούμε τις εξόδους των προηγούμενων βημάτων. Αυτή η συνιστώσα είναι επίσης υπεύθυνη για τον προσδιορισμό των συνθηκών συνένωσης για κάθε παραγόμενο ερώτημα. Το YMAL(*Q*) αποτελείται από την συνένωση των αποτελεσμάτων των παραγόμενων YMAL ερωτημάτων.



Σχήμα 3.1 Αρχιτεκτονική YMAL Συστήματος

3.3. Κανόνες συσχετίσεων

Κατά τη διάρκεια τόσο της τοπικής όσο και της καθολικής ανάλυσης των αποτελεσμάτων ενός ερωτήματος *R*(*Q*), στοχεύουμε στην ανακάλυψη προτύπων που να παρουσιάζουν κάποιο ενδιαφέρον και τα οποία θα τα εκμεταλλευτούμε στη συνέχεια ώστε να συστήσουμε τα YMAL αποτελέσματα. Όπως αναφέρθηκε και πιο

πάνω, τέτοια πρότυπα μπορεί να εντοπιστούν είτε μέσω της προσέγγισης που βασίζεται στο περιεχόμενο του $R(Q)$ είτε μέσω της προσέγγισης που βασίζεται στο σχήμα της D.

Οι *κανόνες συσχετίσεων* (**association rules**) ανακαλύπτουν κρυμμένες συσχετίσεις μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Έτσι, με δεδομένο ένα σύνολο από I αντικείμενα (**items**) κατηγοριοποιημένα με βάση κάποιο χαρακτηριστικό A, όπου $I=\{A.i_1, A.i_2, \ldots, A.i_m\}$ και μία βάση δεδομένων D όπου $A.i_{ij} \in I$ και $A \in D$, ένας κανόνας συσχέτισης είναι ένα επαγωγικό συμπέρασμα της μορφής $X \Rightarrow Y$, όπου $X, Y \subset I$ είναι σύνολα στοιχείων που ονομάζονται *στοιχειοσύνολα* (**itemsets**) και $X \cap Y = \varnothing$.

Ορίζουμε ως *βαθμό υποστήριξης* (**support**) το ποσοστό των δοσοληψιών που περιέχουν ένα στοιχειοσύνολο. Ένα *συχνό στοιχειοσύνολο* (**frequent itemset**) είναι ένα στοιχειοσύνολο του οποίου η υποστήριξη είναι μεγαλύτερη ή ίση από κάποια τιμή κατωφλιού *ελάχιστου βαθμού υποστήριξης* (**minsup**). Ο ελάχιστος βαθμός υποστήριξης βοηθάει στη μείωση των υποψήφιων συχνών συνόλων. Η υποστήριξη ενός στοιχείου -ή ενός συνόλου στοιχείων- για έναν κανόνα συσχέτισης $X \Rightarrow Y$ είναι το ποσοστό των συναλλαγών στη βάση δεδομένων που περιέχουν το $X \cup Y$.

Η πιο κοινή προσέγγιση για την εύρεση των κανόνων συσχέτισης είναι η διάσπαση του προβλήματος σε δύο μέρη: (i) εύρεση συχνών στοιχειοσυνόλων, και (ii) δημιουργία κανόνων από συχνά στοιχειοσύνολα. Αφού έχουν βρεθεί τα συχνά στοιχειοσύνολα, γνωρίζουμε ότι οποιοσδήποτε κανόνας συσχέτισης που παρουσιάζει ενδιαφέρον, $X \Rightarrow Y$,πρέπει να έχει το σύνολο $X \cup Y$ σε αυτό το σύνολο των συχνών στοιχειοσυνόλων. Συνεπώς, το υποσύνολο οποιουδήποτε συχνού στοιχειοσυνόλου είναι επίσης συχνό.

Οι περισσότεροι αλγόριθμοι κανόνων συσχέτισης βασίζονται σε έξυπνους τρόπους για να μειώσουν τον αριθμό των στοιχειοσυνόλων που πρόκειται να μετρηθούν. Αυτά τα πιθανά στοιχειοσύνολα ονομάζονται *υποψήφιοι* (**candidates**), και το σύνολο όλων των καταμετρημένων -πιθανώς συχνών- στοιχειοσυνόλων είναι το *σύνολο των*

υποψήφιων στοιχειοσυνόλων (*C*). Ένα μέτρο της απόδοσης, που χρησιμοποιείται για τους αλγορίθμους των κανόνων συσχέτισης, είναι το μέγεθος του *C*. Όταν έχουν βρεθεί όλα τα συχνά στοιχειοσύνολα, η δημιουργία των κανόνων συσχέτισης είναι μια διαδικασία απλή.

Για την αποδοτική επίλυση του προβλήματος, οι [10] βασίστηκαν σε μια αρχή που ονομάστηκε ***Apriori*** ή περιγραφικά αρχή της προς τα κάτω κλειστότητας. Σύμφωνα με αυτή *αν ένα στοιχειοσύνολο συχνό τότε όλα τα υποσύνολα του είναι επίσης συχνά.* Αυτή η ιδιότητα οδηγεί στην επίλυση του προβλήματος ξεκινώντας από κάτω, δηλαδή από τα συχνά στοιχειοσύνολα ενός αντικειμένου. Με αυτά ως βάση κατασκευάζουμε τα συχνά στοιχειοσύνολα 2 αντικειμένων, μετά 3 κοκ. μέχρι να φτάσουμε σε στοιχειοσύνολα *k*-αντικειμένων που δεν είναι συχνά, και τότε η επαναληπτική διαδικασία σταματά.

Από την διατύπωση του ***Apriori*** και μετά εμφανίστηκαν αρκετές παραλλαγές και βελτιώσεις πάνω στον αλγόριθμο, όπως οι τεχνικές ***hashing*** [PaCY95], τεχνικές ***partitioning*** [SaON95]. Η δικιά μας παραλλαγή, ο αλγόριθμος ΥMALORI, στοχεύει στην επίτευξη της μη-απώλειας πληροφορίας στα συχνά μονοσύνολα αντικειμένων, όπως θα αναλύσουμε εκτενέστερα στη συνέχεια. Επιπλέον, στην πρώτη προσέγγιση στο μοντέλο μας η παραλλαγή του ΥMALORI μας δίνει το ***max-top*** *συχνό* αποτέλεσμα σε κάθε σύνολο αντικειμένων, αγνοώντας τις null τιμές. Κάνουμε χρήση του αλγορίθμου ΥMALORI στην προσέγγιση της τοπικής ανάλυσης, ενώ στην προσέγγιση της καθολικής ανάλυσης χρησιμοποιούμε τον αλγόριθμο ***Apriori***, όπως αυτός διατυπώθηκε από τους [10].

3.4. Συστάσεις Βάσει Περιεχομένου

Στην προσέγγιση των συστάσεων βάσει του περιεχομένου μελετήσαμε τόσο την τοπική όσο και την καθολική ανάλυση. Η τοπική ανάλυση διακρίνει τις πιο συχνές τιμές μέσα στο ***R***(***Q***) και εκμεταλλευόμενη αυτές, μέσω διάφορων τεχνικών, υπολογίζει τις συστάσεις. Από την άλλη πλευρά, η καθολική ανάλυση εντοπίζει τις πιο συσχετισμένες τιμές στη *D*, λαμβάνοντάς και αυτές υπόψη πριν προωθήσει στο χρήστη τις ανάλογες συστάσεις.

3.4.1. Τοπική Ανάλυση

Δεδομένου ενός συνόλου αποτελεσμάτων *R*(*Q*) και των προτύπων που παρουσιάζουν ενδιαφέρον και που προκύπτουν από αυτό, σύμφωνα με κανόνες συσχέτισης, απώτερος στόχος μας είναι να παρουσιάσουμε στο χρήστη ένα σύνολο από YMAL αποτελέσματα με πληθικότητα *m*.

Για τη σύσταση των YMAL ερωτημάτων μέσα στο σύστημά μας, η συνιστώσα της επιλογής χαρακτηριστικών δημιουργεί τα γνωρίσματα του `select` για όλα τα YMAL ερωτήματα, η οποία περιέχει όλα τα χαρακτηριστικά του `B` που εμφανίζονται στις συνθήκες του `P2` του αρχικού ερωτήματος.

Εν συνεχεία, η συνιστώσα της επιλογής σχέσεων, η οποία είναι υπεύθυνη για τα γνωρίσματα του `from` των YMAL ερωτημάτων, επιστρέφει τις σχέσεις `S`, οι οποίες ταυτίζονται με το `R` της αρχικής ερώτησης *Q*, δηλαδή `S` = `R`.

Στο σημείο αυτό, η συνιστώσα επιλογής περιορισμών που δημιουργεί τις συνθήκες επιλογής που περιέχουν τα YMAL ερωτήματα, παρουσιάζει ιδιαίτερο ενδιαφέρον. Στην περίπτωση που εξετάζουμε, δοθέντος ενός ερωτήματος *Q* με χαρακτηριστικά ***A***= {A_1,…,A_m} που εμφανίζονται στις σχέσεις `R`, πρώτα υπολογίζουμε το δυναμικό σύνολο ***A**** που αποτελείται από όλα τα υποσύνολα του ***A***. Έπειτα, ο αλγόριθμος YMALORI δημιουργεί τους περιορισμούς των YMAL ερωτημάτων σύμφωνα με το ***A****.

Ορισμός 1: Δοθέντος ενός ερωτήματος *Q*, του δυναμικού συνόλου ***A**** των χαρακτηριστικών του και το σύνολο των αποτελεσμάτων *R*(*Q*), ο αλγόριθμος YMALORI εξάγει, για κάθε σύνολο χαρακτηριστικών της μορφής {A_i,…,A_j} στο ***A****, μία συνθήκη επιλογής της μορφής:

$$A_i = a_{i_x} \text{ AND } \ldots \text{ AND } A_j = a_{j_y}$$

όπου $a_{i_x} \in \text{dom}(A_i)$, $a_{j_y} \in \text{dom}(A_j)$, και το σύνολο τιμών $\{ a_{i_x}, \ldots, a_{j_y} \}$είναι το πιο συχνά εμφανιζόμενο σύνολο τιμών στο R(Q) για τα {A_i,…,A_j}.

Για κάθε συνθήκη επιλογής που επιστρέφεται από τη λειτουργία επιλογής, δημιουργείται και ένα διαφορετικό YMAL ερώτημα.

- *Ο Αλγόριθμος YMALORI*

Με την παραδοχή ότι συνήθως έχουμε να κάνουμε με τεράστιες βάσεις δεδομένων, οι οποίες περιέχουν μεγάλο αριθμό από διακριτά αντικείμενα οδηγούμαστε στο συμπέρασμα ότι και τα τυχόν σύνολα που απαρτίζονται από τα αντικείμενα αυτά είναι επίσης μεγάλα. Με απώτερο στόχο, λοιπόν, την μείωση των υποψήφιων αντικειμένων μέσα από τόσο μεγάλα σύνολα, οδηγούμαστε στη χρήση της υποστήριξης για την μείωση των υποψήφιων συνόλων, που ορίζει ότι *εάν ένα σύνολο αντικειμένων είναι συχνό τότε όλα τα υποσύνολά του είναι επίσης συχνά* (*ιδιότητα συχνών στοιχειοσυνόλων*). Με βάση την ιδιότητα της αντιθετοαντιστροφής ισχύει και το εξίσου σημαντικό *εάν ένα σύνολο αντικειμένων δεν είναι συχνό τότε και όλα τα υπερσύνολά του είναι επίσης μη συχνά.*

Η στρατηγική αυτή της μείωσης του τεράστιου αριθμού των υποψηφίων συχνών συνόλων λέγεται *απόρριψη βάσει υποστήριξης* και προκύπτει από την ιδιότητα του μεγέθους της υποστήριξης με βάση την οποία, η υποστήριξη ενός συνόλου δεν είναι δυνατόν να υπερβαίνει την υποστήριξη κανενός από τα υποσύνολα του.

Ο αλγόριθμος, που σχεδιάσαμε και χρησιμοποιούμε στο μοντέλο μας, είναι μία παραλλαγή εκείνου των [10]. Ονομάζουμε αυτόν τον αλγόριθμο YMALORI, και στον Πίνακας 3.2 γίνεται μία περιγραφή του αλγορίθμου με χρήση ψευδοκώδικα. Στηριζόμενοι στην ιδιότητα συχνών στοιχειοσυνόλων, εάν γνωρίζουμε ότι ένα στοιχειοσύνολο δεν είναι συχνό, δε χρειάζεται να δημιουργήσουμε κανένα υπερσύνολό του σαν υποψήφιο, επειδή και αυτό αποκλείεται να είναι συχνό. Η βασική ιδέα του αλγορίθμου είναι η δημιουργία υποψηφίων στοιχειοσυνόλων ενός συγκεκριμένου μεγέθους και στη συνέχεια η σάρωση όλων των δεδομένων για να δούμε αν αυτά είναι συχνά.

Κατά τη διάρκεια του k περάσματος, καταμετρούνται τα υποψήφια στοιχειοσύνολα μεγέθους k (C_k). Στην αρχή, για $k = 1$, ο αλγόριθμος κάνει ένα πέρασμα στο σύνολο των I αντικειμένων ώστε να ανιχνεύσει τα μονά συχνά αντικείμενα που αποτελούν το

σύνολο F_1, το σύνολο δηλαδή των αντικειμένων που ικανοποιούν τη συνθήκη της υποστήριξης όπως την έχουμε ορίσει από πριν μέσα στο σύστημα μας. Κάθε στοιχείο μέσα στο I αντιπροσωπεύει μία τιμή ενός χαρακτηριστικού A μέσα στη D, έτσι ώστε $I=\{A.i_1, A.i_2, \ldots, A.i_m\}$, με $A.i_{ij} \in I$ και $A \in D$. Τελικά, μόνο εκείνοι οι υποψήφιοι που είναι συχνοί θα χρησιμοποιηθούν αργότερα για τη δημιουργία υποψηφίων στο επόμενο πέρασμα -κάτι που σημαίνει ότι το F_k χρησιμοποιείται για τη δημιουργία του C_{k+1}.

Αμέσως μετά τον εντοπισμό του συνόλου F_1, και πριν τη δημιουργία του επόμενου συνόλου υποψηφίων k+1, γίνεται ένα «φιλτράρισμα» στα περιεχόμενα του. Στόχος είναι η διατήρηση, μέσα στο σύνολο MF_T, όλων των ***max-top*** *συχνών* τιμών που όμως ανήκουν σε διαφορετικά χαρακτηριστικά. Έτσι, διατηρείται στο MF_T η μέγιστη τιμή που παρατηρείται για κάθε ένα χαρακτηριστικό ξεχωριστά που ανήκει στο I. Για τη δημιουργία υποψηφίων μεγέθους k+1, γίνονται συνενώσεις συχνών στοιχειοσυνόλων που βρίσκονται στο προηγούμενο πέρασμα.

Μία συνάρτηση που ονομάζουμε ***ymalori-gen*** χρησιμοποιείται για τη δημιουργία των υποψήφιων στοιχειοσυνόλων για κάθε πέρασμα μετά το πρώτο (βήμα 18). Όλα τα στοιχειοσύνολα ενός στοιχείου χρησιμοποιούνται σαν υποψήφια για το πρώτο πέρασμα. Εδώ, το σύνολο των συχνών στοιχειοσυνόλων του προηγούμενου περάσματος F_{k-1}, συνενώνεται με τον εαυτό του για να καθορίσει τους υποψηφίους. Μεμονωμένα στοιχειοσύνολα πρέπει να έχουν κοινά όλα τα στοιχεία, εκτός από ένα, έτσι ώστε να συνδυαστούν. Μετά το πρώτο πέρασμα, κάθε συχνό στοιχειοσύνολο συνδυάζεται με όλα τα άλλα συχνά στοιχειοσύνολα. Η ***ymalori-gen*** εγγυάται ότι θα δημιουργήσει ένα υπερσύνολο των συχνών στοιχειοσυνόλων μεγέθους k, $C_k \supset F_k$, όταν η είσοδος είναι F_{k-1}. Η λειτουργία της ***ymalori-gen*** παρουσιάζεται αναλυτικά στα βήματα 35-44.

Τα βήματα 19-25 του αλγορίθμου υπολογίζουν την *υποστήριξη* των υποψηφίων συνόλων. Η συνάρτηση ***subset*** (βήμα 21) βρίσκει όλα τα υποσύνολα k-αντικειμένων που προκύπτουν από την δοσοληψία. Στο βήμα 26 γίνεται ο υπολογισμός των συχνών αντικειμένων που αποτελούν το σύνολο F_k για του διαφορετικούς συνδυασμούς των αντικειμένων με ίδιο k. Έπειτα από τον υπολογισμό των διαφορετικών συνδυασμών,

επαναλαμβάνεται και πάλι το «φιλτράρισμα» με σκοπό αυτή τη φορά να διατηρήσουμε το ***max-top*** *συχνό* αποτέλεσμα για κάθε ένα σύνολο ***k***-στοιχείων (βήματα 27-30), το οποίο και διατηρούμε τελικά (βήμα 31). Ο συνολικός αριθμός επαναλήψεων του εξωτερικού βρόγχου του αλγορίθμου είναι k_{max+1} όπου k_{max} είναι το μέγιστο μήκος που υπάρχει στο σύνολο των συχνών αντικειμένων. Εν τέλη, επιστρέφεται το σύνολο με όλα εκείνα τα ***max-top*** συχνά σύνολα αντικειμένων που έχουμε διατηρήσει (βήμα 33).

Πίνακας 3.2 Αλγόριθμος YMALORI

Είσοδος:Το σύνολο των αποτελεσμάτων του ερωτήματος του χρήστη ***R(Q)***

Έξοδος: Το σύνολο των ***max-top*** *συχνών* αποτελεσμάτων για κάθε σύνολο χαρακτηριστικών. Για τα μονοσύνολα αντικειμένων δίνονται τα ***max-top*** *συχνά* αποτελέσματα για κάθε διαφορετικό χαρακτηριστικό.

```
1: Begin        //αρχικοποίηση μεταβλητών
2: k = 1;  //k-σύνολο από τα συχνά σύνολα αντικειμένων k-στοιχείων
3: MF_i = Ø; // max-top συχνό σύνολο ενός αντικειμένου
4: MFc = Ø; // max-top συχνό σύνολο υποψήφιων k-αντικειμένων
5: MF_T = Ø;  // max-top συχνό σύνολο αντικειμένων

//υπολογισμός συχνών μονοσυνόλων
6: F_k={A.i|A.i ∈ I^σ({i})≥N x minsup}; //I:σύνολο από i αντικείμενα
                        //A:χαρακτηριστικό αντικειμένου i
                       //σ({i}):αριθμός εμφανίσεων αντικειμένου i
7: for all F_k+1 do
8:   if (A = A.previous) then //A.previous:χαρακτηριστικό
                         //προηγούμενου αντικειμένου που ελέγχθηκε
9:      if (σ({i})≥σ({i.previous})) then
          //i.previous: αριθμός εμφανίσεων αντικειμένου που ελέγχθηκε
10:        MF_i = i; //αντικατάσταση στοιχείου
                    //στο σύνολο των max-top συχνών αποτελεσμάτων
11:        if (A ≠ A.previous) then
12:          begin
13:            MF_T = MF_T ∪ MF_i ;
14:            MF_i = i; //προσθήκη στοιχείου στο σύνολο
                         //των max-top συχνών αποτελεσμάτων
15:   end if
16: end for

//υπολογισμός των υποψήφιων συχνών k-συνόλων
17: for (k = 2; F_k-1≠ Ø; k++) do
18:  C_k=ymalori-gen(F_k-1); //δημιουργία & κλάδεμα υποψήφιων k-συνόλων
```

```
19:  for each transaction t in T do
20:     Ct= subset (Ck,t); // υποσύνολο k-αντικειμένων
                           //που προκύπτουν από τη δοσοληψία t
21:     for each candidate itemset c in Ct do
22:        σ(c) = σ(c) + 1;
23:     end for
24:  end for

//υπολογισμός συχνών k-συνόλων
25: Fk={c|c ∈ Ck^σ(c)≥Nxminsup};//c:συνδυασμοί υποψήφιων αντικειμένων
26: for all Fk do
27:   if (σ(c) ≥ σ(c).previous) then
28:     MFc = c; //αντικατάσταση στοιχείου στο σύνολο
                 //των max-top συχνών υποψήφιων αποτελεσμάτων
29: end for
30: MFT=MFT ∪ MFc; //προσθήκη k-συνόλου με μέγιστη συχνότητα εμφάνισης
                   //στο σύνολο των max-top συχνών αποτελεσμάτων
31: return MFT;   //εμφάνιση των max-top συχνών αποτελεσμάτων
32: end.

//η συνάρτηση ymalori-gen() παίρνει ως είσοδο το σύνολο
//των σύνολο συχνών στοιχειοσυνόλων μεγέθους k -1 (Fk-1)
33: function ymalori-gen(F);
34:   begin //αρχικοποίηση μεταβλητών
35:   Ck = ∅;
36:   for each I ∈ Fk-1 do
37:     for each J ≠ I ∈ Fk-1 do
38:      if i-2 of the elements in I and J are equal then
39:        Ck=Ck ∪ {I ∪ J}; //το σύνολο των υποψηφίων μεγέθους k(Ck)
40:     end for
41:   end for
42: end;
```

3.4.2. Καθολική Ανάλυση

Η καθολική ανάλυση στοχεύει στην εκμετάλλευση των ιδιοτήτων της βάσης δεδομένων κατά τη διάρκεια της σύστασης των YMAL αποτελεσμάτων. Σε αυτή την εργασία, η καθολική ανάλυση είναι για εμάς μία διαφορετική προσέγγιση για το μοντέλο μας βάσει περιεχομένου. Θεωρούμε ότι ο υπολογισμός των YMAL αποτελεσμάτων καθοδηγείται τόσο από το σχήμα της *D* όσο και από συγκεκριμένα *στατιστικά* δεδομένα που αφορούν το περιεχόμενο της.

Δοθέντος, λοιπόν, ενός συνόλου αποτελεσμάτων $R(Q)$ και του σχήματος της D, στόχος μας και πάλι είναι να παρουσιάσουμε στο χρήστη ένα σύνολο από YMAL αποτελέσματα πληθικότητας m.

Το σύστημά μας, μέσω της συνιστώσας για την επιλογή των χαρακτηριστικών, χρησιμοποιεί τα γνωρίσματα του `select`, για όλα τα YMAL ερωτήματα, η οποία και εδώ περιέχει όλα τα χαρακτηριστικά του `B` που εμφανίζονται στις συνθήκες του `P2` του αρχικού ερωτήματος.

Στη συνέχεια, η συνιστώσα της επιλογής σχέσεων, η οποία είναι υπεύθυνη για τα γνωρίσματα του `from` των YMAL ερωτημάτων, επιστρέφει τις σχέσεις `S`, τέτοιες ώστε `S=R`.

Τέλος, η συνιστώσα της επιλογής των περιορισμών των YMAL ερωτημάτων, λαμβάνει υπόψη όχι μόνο τα αποτελέσματα του ερωτήματος, αλλά επιπλέον στατιστικά που ήδη διατηρούνται για τα περιεχόμενα της βάσης δεδομένων. Όταν οι τιμές που εμφανίζονται συχνά στο αποτέλεσμα είναι τιμές που είναι σπάνιες στη D αυτό είναι πιο σημαντικό από το όταν αυτές οι τιμές αντιστοιχούν σε τιμές που είναι συχνές στη D. Πιο συγκεκριμένα, για κάθε σύνολο χαρακτηριστικών $\boldsymbol{A^*}$, διατηρούμε το σχετικό σύνολο τιμών $\boldsymbol{V}$ το οποίο εμφανίζεται συχνά ταυτόχρονα στο $R(Q)$ και στη D χρησιμοποιώντας τον τύπο:

$$\frac{συχνότητα\ εμφάνισης(V, R(Q))}{συχνότητα\ εμφάνισης(V, D)} \qquad \text{Εξ. 3.1}$$

όπου η *συχνότητα εμφάνισης* $(V, R(Q))$ συμβολίζει τον αριθμό εμφανίσεων του $\boldsymbol{V}$ στο $R(Q)$ και η *συχνότητα εμφάνισης* (V, D) συμβολίζει τον αριθμό εμφανίσεων του $\boldsymbol{V}$ στη D. Για την εύρεση του συνόλου $\boldsymbol{A^*}$, κάνουμε χρήση του αλγορίθμου *Apriori* [10], όπως αναφέρθηκε και πιο πάνω.

Τελικά, η τιμή του χαρακτηριστικού που ικανοποιεί στο μέγιστο αυτή τη συνθήκη, υπερισχύει των άλλων αντίστοιχων τιμών του ίδιου χαρακτηριστικού, και είναι αυτή που θα χρησιμοποιηθεί στη συνέχεια για τη δημιουργία των YMAL ερωτημάτων.

3.5. Συστάσεις Βάσει Σχήματος

Όσον αφορά τη δημιουργία συστάσεων βάσει σχήματος, προσεγγίσαμε και εδώ τόσο την τοπική όσο και της καθολική ανάλυση. Στην τοπική ανάλυση δημιουργούνται απευθείας ενώσεις με τις συχνότητες των τιμών στο διευρυμένο $R(Q)$. Ενώ, στην καθολική ανάλυση δημιουργούνται απευθείας ενώσεις με τις συσχετίσεις μεταξύ των σχέσεων στη D.

3.5.1. Τοπική Ανάλυση

Στην προσέγγιση της τοπικής ανάλυσης προχωρούμε στη δημιουργία διευρυμένων ερωτημάτων βάσει των συχνών τιμών μέσα στο $R(Q)$. Οι συχνές αυτές τιμές αποτελούν τα ενδιαφέροντα πρότυπα πάνω στα οποία βασιζόμαστε για περαιτέρω επέκταση.

Η συνιστώσα για την επιλογή των χαρακτηριστικών, χρησιμοποιεί ειδικά γνωρίσματα για το `select`. Δημιουργεί αυτά τα γνωρίσματα βασιζόμενη στα χαρακτηριστικά που εμφανίζονται στις σχέσεις `S` των γνωρισμάτων του `from` του υπό κατασκευή YMAL ερωτήματος, αγνοώντας εκείνα τα χαρακτηριστικά που περιέχουν μη ενδιαφέρουσα πληροφορία, όπως για παράδειγμα τιμές πρωτευόντων και ξένων κλειδιών που περιέχουν κωδικοποιημένα δεδομένα. Ακόμη και στην περίπτωση που οι σχέσεις της `S` καταλήγουν σε εξαίρετες περιπτώσεις να ταυτίζονται με αυτές της `R`, για παράδειγμα στην περίπτωση που το `R` περιείχε όλους τους πίνακες της D, το `B` θα ήταν αρκετά τροποποιημένο από το `B` των προηγούμενων προσεγγίσεων. Σε μία τέτοια περίπτωση, η συνιστώσα για την επιλογή των χαρακτηριστικών θα πρόσθετε επιπλέον στο `B` όλα εκείνα τα ‘χρήσιμα’ πεδία που αναφέρονται στα γνωρίσματα του `from` του υπό κατασκευή YMAL ερωτήματος, και τα οποία δεν ταυτίζονται με εκείνα που περιέχονταν μέσα στο `A` της Q.

Η συνιστώσα της επιλογής σχέσεων, υπεύθυνη για τα γνωρίσματα του `from` των YMAL ερωτημάτων, επιστρέφει τις σχέσεις S, οι οποίες είναι ένα υπερσύνολο του R. Για να υπολογίσουμε το S, αρχικά δημιουργούμε νέα ερωτήματα της μορφής:

`select count(*) from` $R_{candidate}$, (Q) `T where` R_i`.id=T.id;`

όπου $R_{candidate}$ η υποψήφια σχέση που θα προστεθεί στο S. Ως $R_{candidate}$ επιλέγουμε τις k πιο σχετιζόμενες σχέσεις από εκείνες που υπάρχουν στο R, όπου το k καθορίζει τον περιορισμό της πληθικότητας. Για παράδειγμα, για k =1, ως $R_{candidate}$ θα λάβουμε υπόψη ως προς τον έλεγχο σχέσεις που βρίσκονται σε απόσταση ίση με ένα από τις σχέσεις που αναφέρονται στο R.

Η κάθε σχέση $R_{candidate}$ συσχετίζεται με τη σχέση R_i, υποσύνολο των σχέσεων που υπάρχουν μέσα στο Q, με ένα βαθμό p. Ορίζουμε ως *βαθμό* $p(R_{candidate}, R_i)$ για τις σχέσεις $R_{candidate}$ και R_i μία ποσότητα ανάλογη του μεγέθους του αποτελέσματος της συνένωσης των δύο σχέσεων προς το μέγεθος του $R_{candidate}$ επί το μέγεθος του R_i, σύμφωνα με τον τύπο:

$$p(\mathrm{R_{cadidate}}, \mathrm{R_i}) = \frac{size(\mathrm{R_{cadidate}} \bowtie \mathrm{R_i})}{size(\mathrm{R_{cadidate}}) * \mathrm{size(R_i)}} \qquad \text{Εξ. 3.2}$$

Τελικά, θα επιλεγεί μόνο μία $R_{candidate}$ σχέση, και θα είναι εκείνη της οποίας ο βαθμός $p(R_{candidate}, R_i)$ είναι ο μεγαλύτερος όλων όσων ελέγχθηκαν. Με αυτόν τον τρόπο, η συνιστώσα της επιλογής σχέσεων, κατασκευάζει το S προσθέτοντας στο R την $R_{candidate\_max}$ σχέση.

Η συνιστώσα της επιλογής των περιορισμών των YMAL ερωτημάτων, μέσω του αλγορίθμου YMALORI δημιουργεί τους περιορισμούς των YMAL ερωτημάτων σύμφωνα με το ***A****.

3.5.2. Καθολική Ανάλυση

Μέσω της καθολικής ανάλυσης εκμεταλλευόμαστε το σχήμα της βάσης δεδομένων κατά τη διάρκεια της σύστασης των YMAL αποτελεσμάτων. Στόχος μας είναι η δημιουργία διευρυμένων ερωτημάτων βάσει του σχήματος της *D*, και η παρουσίαση στο χρήστη ενός συνόλου από YMAL αποτελέσματα πληθικότητας *m*.

Όμοια με την τοπική ανάλυση της εν λόγω προσέγγισης, η συνιστώσα για την επιλογή των χαρακτηριστικών, χρησιμοποιεί ειδικά γνωρίσματα για το `select`, βασιζόμενη στα χαρακτηριστικά που εμφανίζονται στις σχέσεις `S` των γνωρισμάτων του `from` του υπό κατασκευή YMAL ερωτήματος, αγνοώντας χαρακτηριστικά με μη ενδιαφέρουσα πληροφορία. Και πάλι αναφερόμαστε σε ένα `B` το οποίο περιέχει όλα εκείνα τα 'χρήσιμα' πεδία που αναφέρονται στα γνωρίσματα του `from` του υπό κατασκευή YMAL ερωτήματος, και τα οποία δεν ταυτίζονται με εκείνα που περιέχονταν μέσα στο `A` της *Q*.

Η συνιστώσα της επιλογής σχέσεων, επιστρέφει τις σχέσεις `S`, οι οποίες είναι ένα υπερσύνολο του `R`. Για να υπολογίσουμε το `S`, διατηρούμε τις συσχετίσεις μεταξύ των σχέσεων στην *D*. Μία σχέση R_i συσχετίζεται με τη σχέση R_j με ένα βαθμό *p*. Ορίζουμε ως *βαθμό* $p(R_i, R_j)$ για τις σχέσεις R_i και R_j μία ποσότητα ανάλογη του μεγέθους του αποτελέσματος της συνένωσης των δύο σχέσεων προς το μέγεθος του R_i επί το μέγεθος του R_j, σύμφωνα με τον τύπο:

$$p(\mathrm{R_i}, \mathrm{R_j}) = \frac{size(\mathrm{R_i} \bowtie \mathrm{R_j})}{size(\mathrm{R_i}) * \mathrm{size}(\mathrm{R_j})} \qquad \text{Εξ. 3.3}$$

Ο βαθμός *p* είναι μία ποσότητα προϋπολογισμένη, την οποία χρησιμοποιούμε δυναμικά για κάθε επέκταση των R_k σχέσεων. Με αυτόν τον τρόπο, κατασκευάζουμε το `S` προσθέτοντας στο `R` τις *k* πιο σχετιζόμενες σχέσεις από εκείνες στο `R`, όπου το *k* καθορίζεται από μία τιμή εισόδου για το περιορισμό της πληθικότητας. Όσο πιο μεγάλος είναι ο βαθμός *p* μιας σχέσης, τόσο πιο συσχετιζόμενη θεωρείται η σχέση.

Στο τελικό στάδιο, η συνιστώσα της επιλογής των περιορισμών των YMAL ερωτημάτων, ομοίως με την τοπική ανάλυση, μέσω του αλγορίθμου YMALORI δημιουργεί τους περιορισμούς των YMAL ερωτημάτων σύμφωνα με το ***A****.

ΚΕΦΑΛΑΙΟ 4. ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΥ

Στο κεφάλαιο αυτό γίνεται περιγραφή της υλοποίησης της μεθόδου συστάσεων που περιγράφτηκε στα προηγούμενο κεφάλαιο. Παράλληλα, αναλύονται τα αποτελέσματα που ανάγονται από την υλοποίηση αυτή, και γίνεται μελέτη των προσεγγίσεων ξεχωριστά με τους παράγοντες που τις επηρεάζουν. Τέλος, μέσω της σύγκρισης των πειραματικών αποτελεσμάτων, παρουσιάζουμε μία πλήρη εικόνα των διαφόρων προσεγγίσεων, εξάγοντας χρήσιμα συμπεράσματα.

4.1. Περιγραφή υλοποίησης

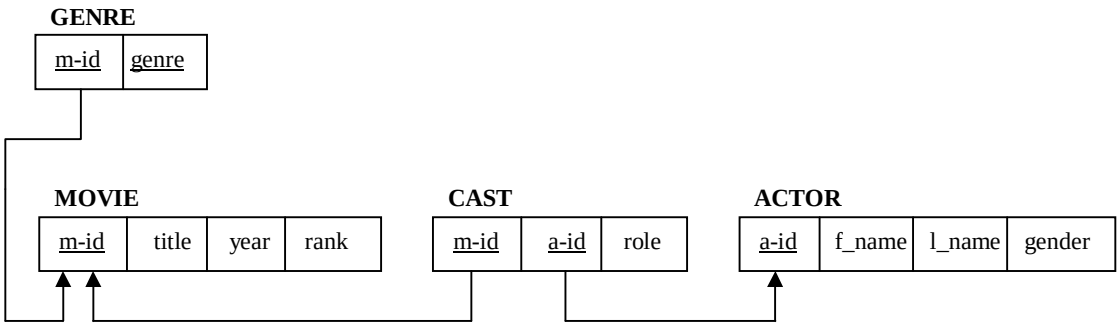
Στην ενότητα αυτή παρουσιάζουμε την υλοποίηση που δημιουργήθηκε για το σκοπό της εργασίας μας. Η υλοποίηση αποτελεί μία αναπαράσταση ενός συστήματος συστάσεων και λειτουργεί βάσει τεσσάρων διαφορετικών προσεγγίσεων. Οι προσεγγίσεις αυτές είναι σε σχέση με συστάσεις που γίνονται βάσει περιεχομένου σε επίπεδο τοπικής και καθολικής ανάλυσης, καθώς και βάσει σχήματος, και πάλι, σε επίπεδο τοπικής και καθολικής ανάλυσης.

Για τον εντοπισμό του συνόλου των αποτελεσμάτων προς σύσταση, δεδομένου ερωτημάτων από την πλευρά των χρηστών, χρησιμοποιήσαμε την αντικειμενοστραφή γλώσσα προγραμματισμού Java, πάνω σε βάση δεδομένων υλοποιημένη με MySQL. Η υλοποίηση του συστήματος πάνω σε ένα ήδη υπάρχον σύστημα βάσης δεδομένων έχει πολλά πλεονεκτήματα, όπως η μεταφερσιμότητα και η ευκολία υλοποίησης. Το σύστημά μας μπορεί εύκολα να προσπελαστεί από έναν απλό φυλλομετρητή

κάνοντας χρήση ενός γραφικού περιβάλλοντος που παρέχει μεγάλη αλληλεπίδραση με το χρήστη. Το περιβάλλον διεπαφής υλοποιήθηκε με την τεχνολογία της JSP.

4.1.1. Δεδομένα

Η βάση δεδομένων που χρησιμοποιήσαμε περιέχει ταινίες και προέρχεται από την IMDB [17], λεπτομέρειες για το μέγεθος και τον τύπο των δεδομένων φαίνονται στον Πίνακα , ενώ το σχήμα της παρουσιάζεται στο Σχήμα 4.1.



Σχήμα 4.1 Σχήμα Βάσης Δεδομένων

Πίνακας 4.1 Δομή Βάσης Δεδομένων

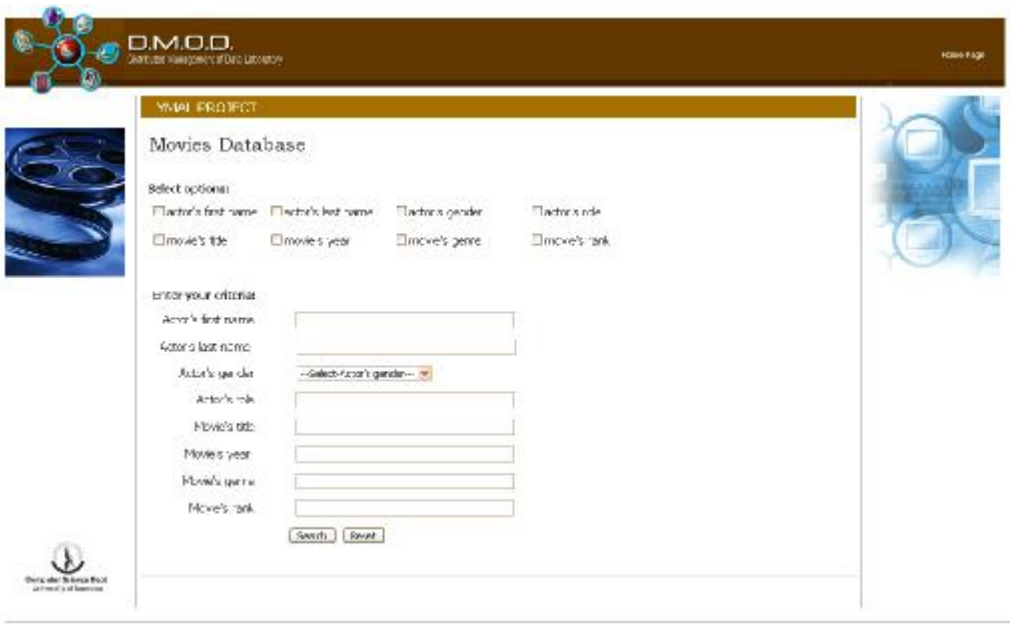
ΠΙΝΑΚΑΣ	ΕΓΓΡΑΦΕΣ	ΤΥΠΟΣ	ΜΕΓΕΘΟΣ
Cast	~1,391,960	InnoDB	176,4 MB
Genre	~96,431	InnoDB	10,5 MB
Movies	~70,152	InnoDB	16,6 MB
Actor	~475,890	InnoDB	67,7 MB
Σύνολο	~2,034,433	InnoDB	271,2 MB

4.1.2. Βασική Λειτουργία

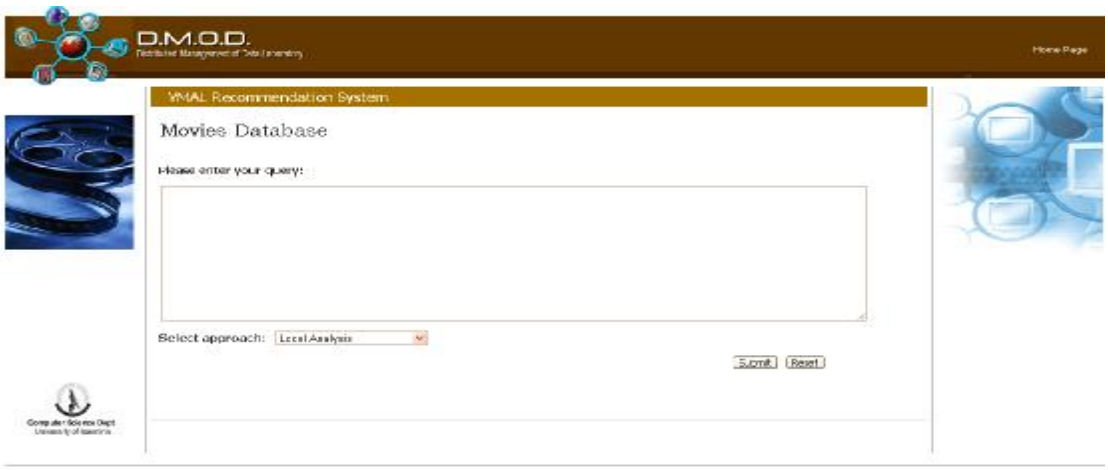
Ο χρήστης μπορεί να υποβάλλει τα ερωτήματά του μέσω της γλώσσας SQL (Σχήμα 4.3) ή κάνοντας χρήση διαθέσιμων φορμών που παρέχει το σύστημα (Σχήμα 4.2). Έπειτα από την εκτέλεση του ερωτήματος, παρουσιάζεται στο χρήστη τόσο το σύνολο του αποτελέσματος του αρχικού τους ερωτήματος, όσο και ένα σύνολο προτεινόμενων αποτελεσμάτων, των YMAL συστάσεων (Σχήμα 4.4). Ακόμη, μέσω τους συστήματος, παρουσιάζεται μία επεξήγηση για κάθε μία από τις YMAL

συστάσεις, δηλαδή το πώς η συγκεκριμένη σύσταση σχετίζεται με το πραγματικό αποτέλεσμα του αρχικού ερωτήματος. Αρχικά, παρουσιάζουμε μία σύσταση που αντιπροσωπεύει κάθε YMAL ερώτημα. Αν ο χρήστης το επιλέξει, εμφανίζονται επιπλέον συστάσεις για το εκάστοτε YMAL ερώτημα.

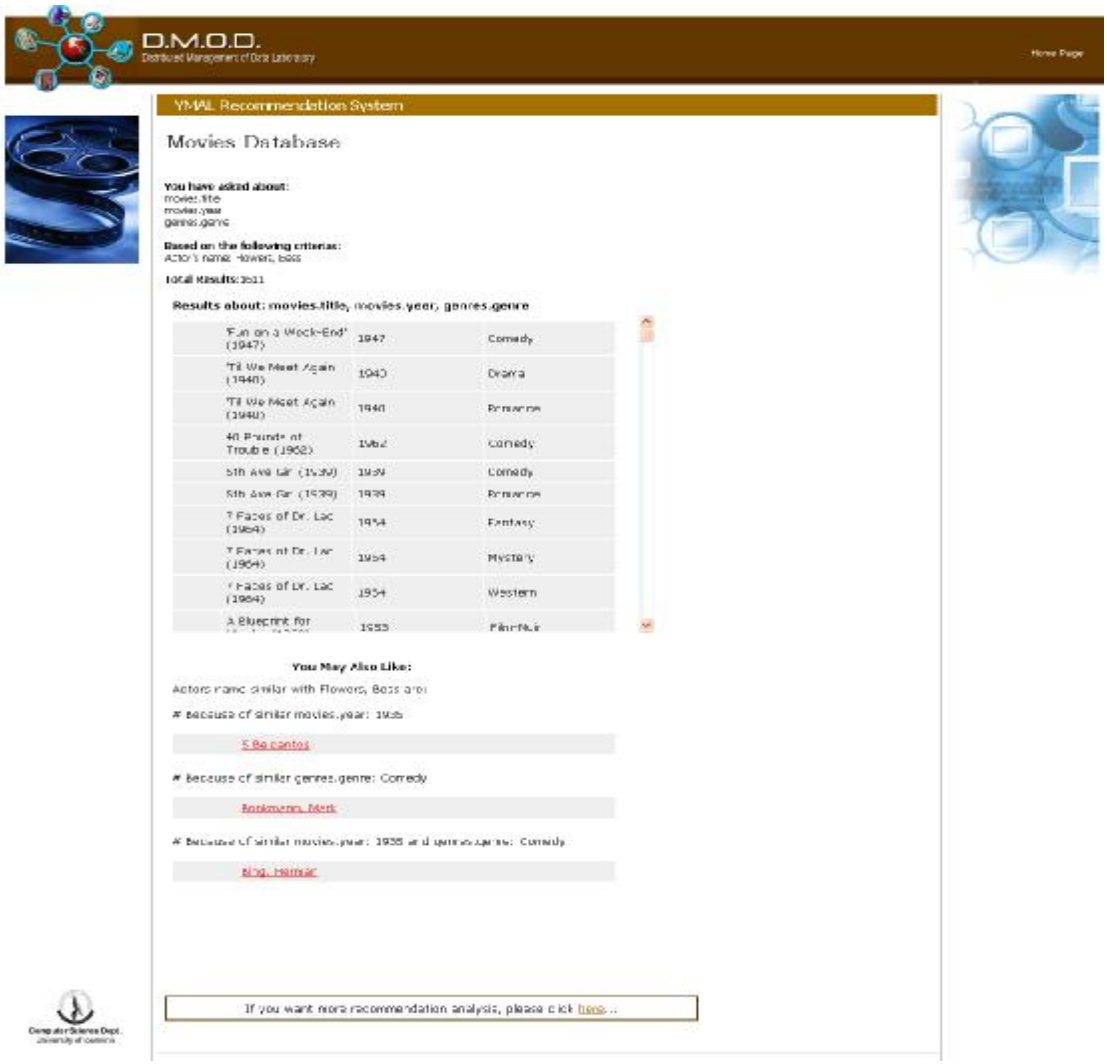
Στην περίπτωση που ο χρήστη κρίνει ενδιαφέρουσες τις YMAL συστάσεις και θα ήθελε να εμβαθύνει ακόμη περισσότερο, δίνεται η ευκαιρία για περαιτέρω επεξεργασία της αρχικής του ερώτησης. Ο χρήστης επιλέγει τον κατάλληλο σύνδεσμο και το σύστημα μειώνει δυναμικά το βαθμό της ελάχιστης υποστήριξης εντοπίζοντας περισσότερα συχνά στοιχειοσύνολα. Εν συνεχεία, εκτελούνται YMAL ερωτήματα βάσει των νέων συχνών στοιχειοσυνόλων και τα αποτελέσματα παρουσιάζονται ως YMAL συστάσεις στο χρήστη σε ένα νέο παράθυρο (Σχήμα 4.5).



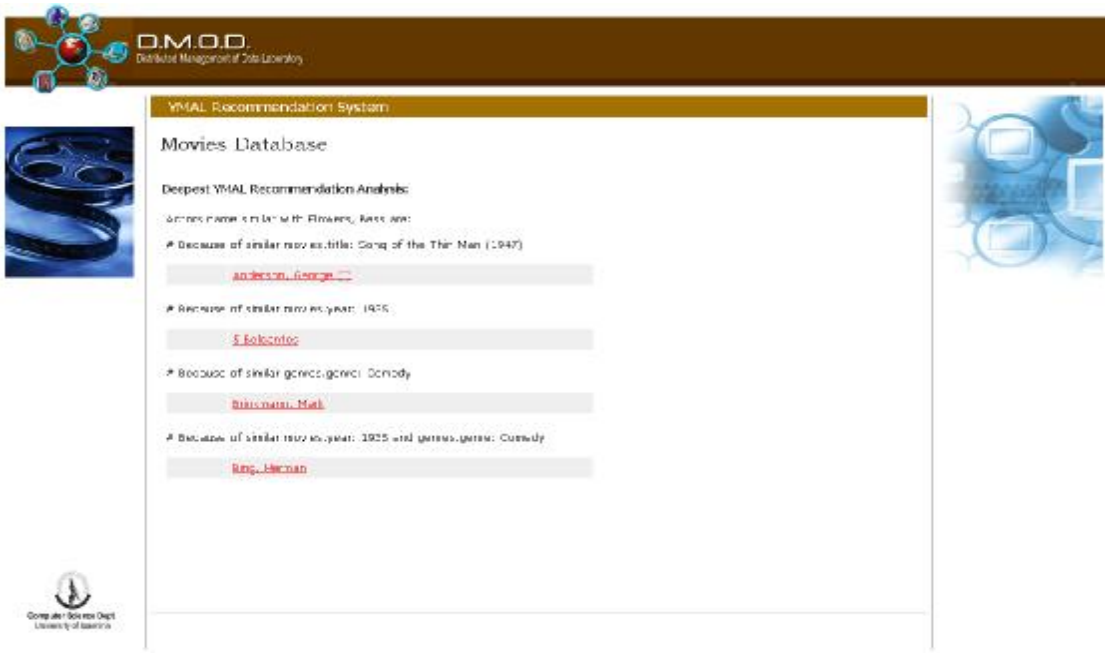
Σχήμα 4.2 Το Περιβάλλον Διεπαφής του YMAL Συστήματος Συστάσεων: Διατύπωση Αρχικού Ερωτήματος Μέσω Φόρμας



Σχήμα 4.3 Το Περιβάλλον Διεπαφής του YMAL Συστήματος Συστάσεων: Διατύπωση Αρχικού Ερωτήματος Με Χρήση SQL



Σχήμα 4.4 Αποτελέσματα Ερώτησης Χρήστη και YMAL Συστάσεις



Σχήμα 4.5 Επιπλέον YMAL Συστάσεις: Με Δυναμική Αλλαγή της Υποστήριξης

4.2. Πειραματικά Αποτελέσματα

Η *αποτελεσματικότητα* των συστάσεων, θα μπορούσε να καταγραφεί, σύμφωνα με τη βιβλιογραφία [2], κάνοντας μετρήσεις στη *κάλυψης* (*coverage*) των αλγορίθμων συστάσεων. Η μετρική της κάλυψης αναφέρεται στα ποσοστά των αντικειμένων για τα οποία ένα σύστημα συστάσεων είναι ικανό να κάνει προβλέψεις. Στην παρούσα έρευνα θα μετρήσουμε την αποτελεσματικότητα του αλγορίθμου μας στις τέσσερις διαφορετικές προσεγγίσεις δίνοντας έμφαση στο *βαθμό κάλυψης*, στην *ομοιομορφία κάλυψης*, καθώς και στη *διαφοροποίηση* (*diversification*) [13].

Έστω M ο μέγιστος αριθμός διαφορετικών συνόλων στο $\boldsymbol{A^*}$, και M' τα διαφορετικά σύνολα του $\boldsymbol{A^*}$, υποσύνολα του M, τα οποία προκύπτουν μέσω της συνιστώσας επιλογής περιορισμών του YMAL συστήματος συστάσεων.

Ορισμός 2: Ονομάζουμε *βαθμό κάλυψης* (*coverage degree*) C_d το ποσοστό των όμοιων συνόλων του M' σε σχέση με εκείνα του M.

Μέσω της συνιστώσας για τη δημιουργία των YMAL ερωτημάτων, καθορίζεται δυναμικά το *ιδανικό μέγεθος* m των αποτελεσμάτων των διαφορετικών συνόλων του M' που αναμένεται να προκύψουν. Δοθείσας μίας ερώτησης Q έστω ότι προκύπτουν M' διαφορετικά σύνολα του $\boldsymbol{A^*}$, με το καθένα από αυτά να έχει μέγεθος m_i.

Ορισμός 3: Ονομάζουμε *ομοιομορφία κάλυψης* (***uniform coverage***) C_u το ποσοστό του αθροίσματος του πραγματικού μεγέθους των M' διαφορετικών συνόλων του $\boldsymbol{A^*}$, όπως προκύπτουν μέσω της συνιστώσας για τη δημιουργία των YMAL ερωτημάτων, προς το άθροισμα του ιδανικού μεγέθους των M' διαφορετικών συνόλων του $\boldsymbol{A^*}$, σύμφωνα με τον τύπο:

$$C_u = \frac{\sum_{i=1}^{M'} m_i}{\sum_{i=1}^{M'} m}$$

Θεωρούμε *ιδανικό σύνολο* d YMAL συστάσεων, αποτελέσματα που, ανεξαρτήτως των διαφορετικών συνόλων του M' και του μεγέθους m, δεν επαναλαμβάνονται. Δοθείσας μίας ερώτησης Q έστω ότι προκύπτουν M' διαφορετικά σύνολα του $\boldsymbol{A^*}$, τα οποία έχουν d_i πλήθος διαφορετικών YMAL συστάσεων.

Ορισμός 4: Καλούμε *διαφοροποίηση* (***diversification***) Rec_d την ανομοιομορφία των YMAL συστάσεων, και την υπολογίζουμε ως το ποσοστό του πλήθους των πραγματικά διαφορετικών YMAL συστάσεων προς το πλήθος των ιδανικά διαφορετικών YMAL συστάσεων, σύμφωνα με τον τύπο:

$$Rec_d = \frac{\sum_{i=1}^{M'} d_i}{\sum_{i=1}^{M'} d}$$

Επιπλέον, στη συνέχεια θα δείξουμε πως ο βαθμός *επιλεκτικότητας* (***selectivity***) κάθε ερώτησης παίζει ξεχωριστό ρόλο στην ανάλυση ευαισθησίας (***sensitivity analysis***) των αποτελεσμάτων.

Η *ανάλυση ευαισθησίας* (***sensitivity analysis***) [14], είναι η μελέτη της διατήρησης μίας βέλτιστης λύσης υπό το φως μεταβολής των αριθμητικών δεδομένων του μοντέλου

του προβλήματος. Ουσιαστικά εδώ προσδιορίζονται τα διαστήματα τιμών κάποιων συντελεστών για τα οποία η βέλτιστη λύση διατηρείται.

Σε αυτή την ενότητα, παρουσιάζονται τα αποτελέσματα της υλοποίησής μας όπως αυτή αναλύθηκε διεξοδικά πιο πάνω. Στον Πίνακα 4.2 δίνονται όλες οι παράμετροι του συστήματος, ώστε ο αναγνώστης να έχει μία πιο σαφή εικόνα για το πώς αυτές επηρεάζουν το σύνολο των YMAL αποτελεσμάτων.

Πίνακας 4.2 Παράμετροι YMAL Συστήματος Συστάσεων

ΠΑΡΑΜΕΤΡΟΙ ΣΥΣΤΗΜΑΤΟΣ	
minsup	υποστήριξη
(V, R(Q))	συχνότητα εμφάνισης στο αποτέλεσμα
(V, D)	συχνότητα εμφάνισης στη βάση δεδομένων
p	βαθμός συσχέτισης
k	πιο σχετιζόμενες σχέσεις στο R
selectivity	επιλεκτικότητα

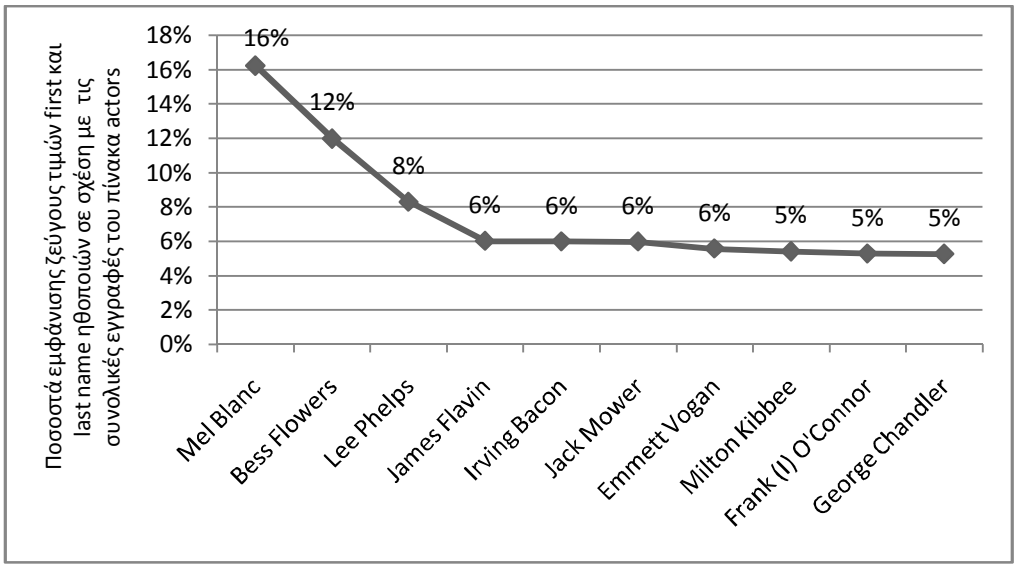
Στις υποενότητες που ακολουθούν, παρουσιάζονται διάφορα πειράματα που έγιναν και τα αποτελέσματα που προέκυψαν τροποποιώντας τις τιμές των διαφόρων παραμέτρων του συστήματος, σε κάθε μία από τις τέσσερις προσεγγίσεις που υλοποιήθηκαν.

4.2.1. Δημιουργία Ερωτήσεων

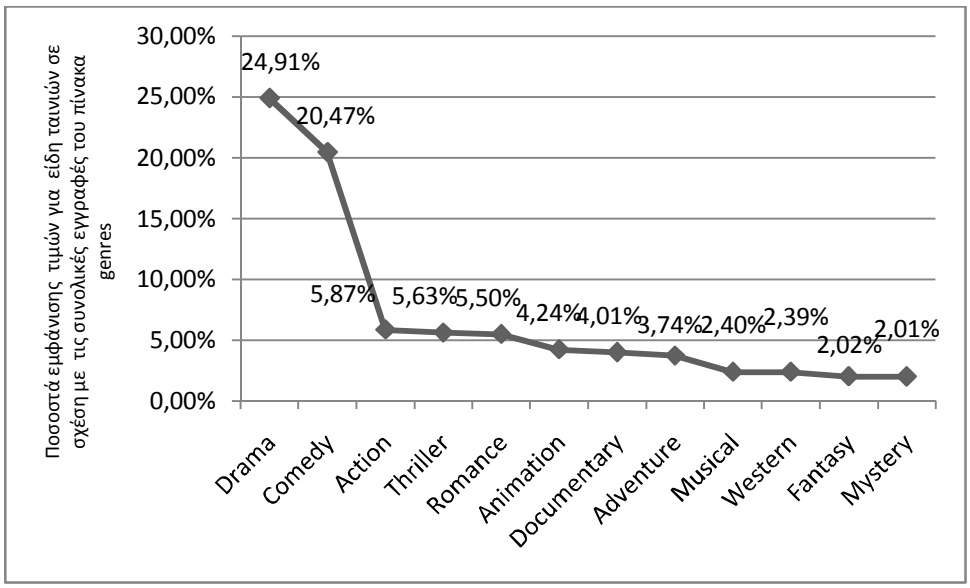
Το Σχήμα 4.6 παρουσιάζει τις πιο δημοφιλείς εγγραφές στη βάση δεδομένων για τα ονόματα των ηθοποιών από τον πίνακα Actor, το Σχήμα 4.7 παρουσιάζει τις πιο δημοφιλείς εγγραφές στη βάση δεδομένων για τα είδη των ταινιών από τον πίνακα Genre, και τέλος το Σχήμα 4.8 παρουσιάζει τις πιο δημοφιλείς εγγραφές στη βάση δεδομένων για τα έτη των ταινιών από τον πίνακα Movies.

Με βάση αυτά τα δεδομένα θα δούμε πώς μεταβάλλονται τα αποτελέσματά μας εάν η αρχική ερώτηση του χρήστη περιέχει μία από αυτές τις εγγραφές. Θα παρουσιάσουμε κοινές ερωτήσεις για κάθε μία προσέγγιση χρησιμοποιώντας ως βαθμούς υποστήριξης τους *5, 15* και *25*, για τις δημοφιλείς εγγραφές της βάσης δεδομένων. Ο

λόγος που επιλέγουμε δημοφιλείς εγγραφές από πίνακες, σε αυτό το σημείο των μετρήσεων μας, είναι για να είμαστε πιο ευέλικτοι στις μεταβολές της υποστήριξης μέσα από μία *μεγαλύτερη κλίμακα τιμών*, με σκοπό να πετυχαίνουμε καλύτερο βαθμό κάλυψης καθώς και ομοιομορφία κάλυψης, μελετώντας παράλληλα πως τροποποιείται η διαφοροποίηση.



Σχήμα 4.6 Δημοφιλείς Ηθοποιοί



Σχήμα 4.7 Δημοφιλή Είδη Ταινιών

Έστω η αρχική ερώτηση του χρήστη u_1 είναι η $Q1$:

```
select  CAST.role, GENRE.genre
from    ACTOR, CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        ACTOR.pid = CAST.pid and
        ACTOR.fname= 'Lee' and
        ACTOR.lname= 'Phelps';
```

Το $R(Q1)$ πλέον θα είναι η είσοδος στο YMAL σύστημα συστάσεων. Βάσει της $Q1$ δημιουργήσαμε άλλες *5* παρόμοιες ερωτήσεις αλλάζοντας μόνο το όνομα του ηθοποιού, χρησιμοποιώντας συνολικά τους πέντε πρώτους πιο δημοφιλείς της βάσης δεδομένων σύμφωνα με το Σχήμα 4.6.

Και έστω, η αρχική ερώτηση του χρήστη u_2 είναι η $Q2$:

```
select  CAST.role, MOVIES.year
from    CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        GENRE.genre= 'Western'
```

Ομοίως, το $R(Q2)$ θα είναι η είσοδος στο YMAL σύστημα συστάσεων. Βάσει της $Q2$ δημιουργήσαμε και πάλι άλλες *5* παρόμοιες ερωτήσεις αλλάζοντας το είδος της ταινίας, επιλέγοντας δημοφιλή είδη της βάσης δεδομένων σύμφωνα με το Σχήμα 4.7.

Τέλος, για πιο αποδοτικές μετρήσεις στις δύο προσεγγίσεις των συστάσεων βάσει σχήματος, χρησιμοποιήσαμε επιπλέον το μοτίβο της ερώτησης $Q3$:

```
select  CAST.role
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        MOVIES.year = '2003';
```

καθώς και το μοτίβο της ερώτησης $Q4$:

```
select  MOVIES.year
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        CAST.role = 'Policeman';
```

Βάσει της *Q*3 και της *Q*4 δημιουργήσαμε ομοίως άλλες *5* παρόμοιες ερωτήσεις γα κάθε μοτίβο ερώτησης, αλλάζοντας στην πρώτη περίπτωση τη χρονιά της ταινίας, και στη δεύτερη περίπτωση το ρόλο των ηθοποιών, επιλέγοντας από τυχαίες εγγραφές της βάσης δεδομένων και διατηρώντας σταθερό το βαθμό της υποστήριξης στο *5*.

Για κάθε μία προσέγγισης στη συνέχεια, θα περιγράψουμε αρχικά τον τρόπο που δουλέψαμε και θα παρουσιάσουμε εν συνεχεία αντίστοιχα διαγράμματα για τα συνολικά τα δεδομένα που λάβαμε.

4.2.2. Δυναμικός υπολογισμός υποστήριξης

Στο σημείο αυτό θα δούμε πως τροποποιούνται τα αποτελέσματα των συστάσεων για κάθε μία διαφορετική προσέγγιση με βάση τόσο την μεταβολή της υποστήριξης, όσο και των χαρακτηριστικών που περιέχονται στην αρχική ερώτηση του χρήστη.

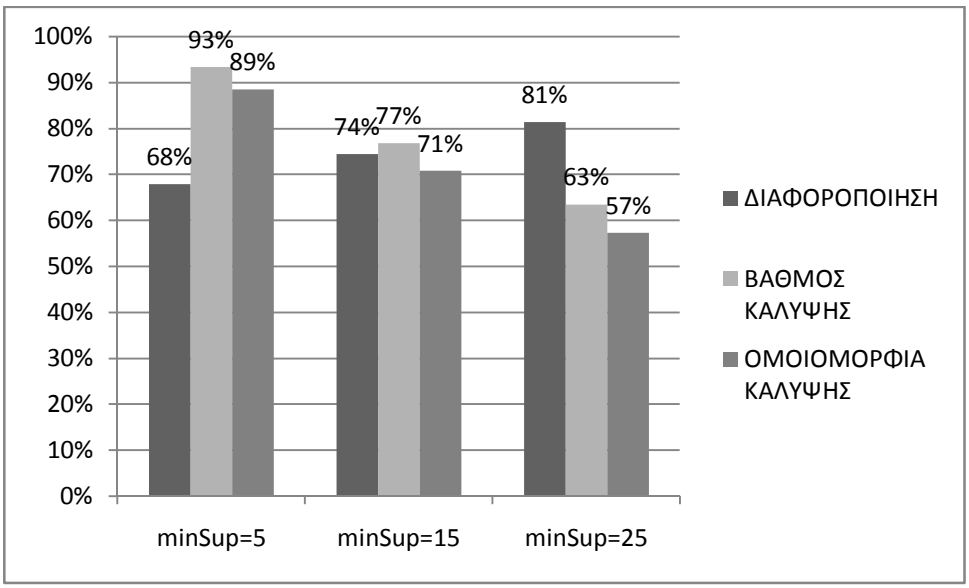
- *Συστάσεις Βάσει Περιεχομένου: Τοπική Ανάλυση*

Για την προσέγγιση της τοπικής ανάλυσης βάσει περιεχομένου επισημαίνουμε ότι ο αλγόριθμος YMALORI μας επιστρέφει ότι μόνο τα ***top-1*** *συχνά σύνολα τιμών* που ανήκουν στο αποτέλεσμα της αρχικής ερώτησης *Q*. Αυτό σημαίνει ότι αυξάνοντας το βαθμό υποστήριξης απλά μειώνουμε τις πιθανότητες να πετύχουμε πλήρη κάλυψη, μειώνοντας τα μεγαλύτερα σύνολα τιμών που επιστρέφει ο αλγόριθμος. Από την άλλη πλευρά, στοχεύουμε στη διαφοροποίηση των αποτελεσμάτων που παρουσιάζουμε στο χρήστη.

Από το Σχήμα 4.8 παρατηρούμε ότι για μικρές τιμές της υποστήριξης πετυχαίνουμε μεγάλα ποσοστά ομοιομορφίας κάλυψης. Αυτό σημαίνει ότι λαμβάνουμε σχεδόν το επιθυμητό πλήθος συστάσεων για τον κάθε πιθανό συνδυασμό αποτελέσματος που προκύπτει μέσα από τον αλγόριθμο YMALORI. Αντίθετα, όσον αφορά τη διαφοροποίηση των αποτελεσμάτων παρατηρούμε ότι οι συστάσεις

επαναλαμβάνονται, δηλαδή ο χρήστης βλέπει επαναληπτικά ίδιες συστάσεις λόγω των διαφορετικών συνδυασμών του αποτελέσματος του YMALORI. Ακόμη, παρατηρείται μία αντιστροφή αυτών των συχνοτήτων, όταν η υποστήριξη λαμβάνει μεγάλες τιμές. Τότε, οι συστάσεις είναι πιο διαφοροποιημένες, αλλά υστερεί η ομοιομορφία κάλυψης, όπως επίσης και ο βαθμός κάλυψης, δηλαδή η δημιουργία συστάσεων για κάθε πιθανό συνδυασμό.

Το συμπέρασμα που ανάγεται μέσα από τα πειραματικά αποτελέσματα για αυτή την προσέγγιση είναι ότι για συχνές εγγραφές στη βάση, εγγραφές με μεγάλη επιλεκτικότητα, οι οποίες αναφέρονται μέσα στη *Q*, επιστρέφει καλύτερα αποτελέσματα με μία ενδιάμεση τιμή υποστήριξης. Η ανάλυση ευαισθησίας μας οδήγησε στο βαθμός υποστήριξης *15*. Με χρήση ενδιάμεσης τιμής υποστήριξης, η ποσοστιαία διαφορά μεταξύ βαθμού κάλυψης, ομοιομορφίας κάλυψης και διαφοροποίησης μειώνεται αρκετά.



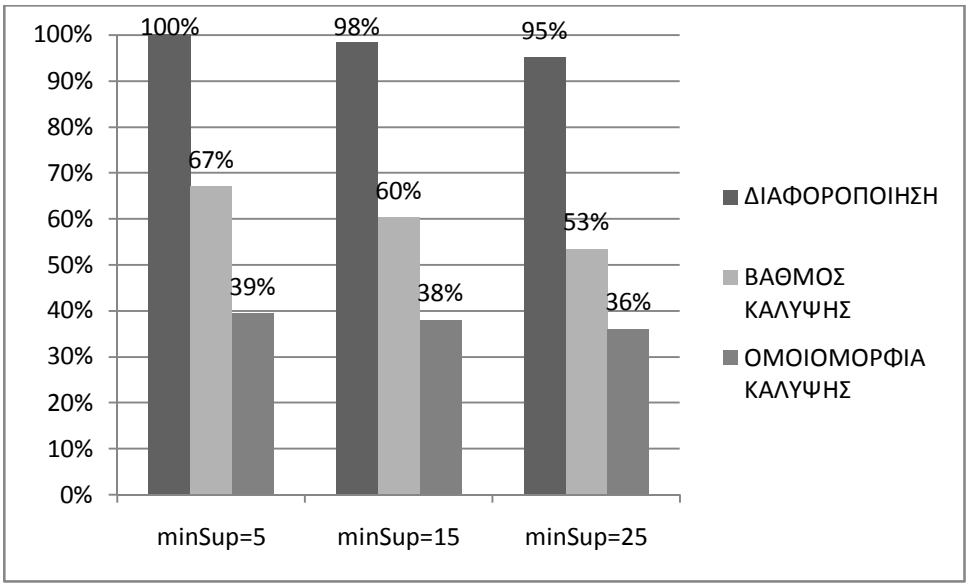
Σχήμα 4.8 Συστάσεις Βάσει Περιεχομένου (Τοπική Ανάλυση): Συσχετίσεις Διαφοροποίησης Και Κάλυψης Των Συστάσεων Λόγω Μεταβολών Της Υποστήριξης

- *Συστάσεις Βάσει Περιεχομένου: Καθολική Ανάλυση*

Στην προσέγγιση της καθολικής ανάλυσης βάσει περιεχομένου, ο αλγόριθμος YMALORI μας επιστρέφει και πάλι τα *top-1 συχνά σύνολα τιμών*, εντοπίζοντάς τα όμως με διαφορετικό τρόπο. Όπως αναφέρθηκε και στην περιγραφή του μοντέλου

που ακολουθήσαμε, στο προηγούμενο κεφάλαιο, κάθε *top-1* συχνό σύνολο τιμών υπολογίζεται ως η μέγιστη τιμή που παίρνει το κλάσμα *συχνότητα εμφάνισης αποτελέσματος στο σύνολο αποτελεσμάτων της Q ÷ συχνότητα εμφάνισης αποτελέσματος στη D*. Αυτό σημαίνει ότι αυξάνοντας το βαθμό υποστήριξης επιστρέφονται συνεχώς νέα αποτελέσματα μέσα από τον YMALORI. Τα νέα αυτά αποτελέσματα έχουν σαν άμεσο αποτέλεσμα τη δημιουργία διαφορετικών κάθε φορά YMAL ερωτήσεων.

Στο Σχήμα 4.9 παρατηρούμε ότι παρότι ο βαθμός υποστήριξης αυξομειώνεται δυναμικά, πετυχαίνουμε πολύ μεγάλα ποσοστά διαφοροποίησης των συστάσεων. Αντίθετα, όσον αφορά την ομοιομορφία κάλυψης των αποτελεσμάτων παρατηρούμε ότι η συγκεκριμένη μέθοδος δεν είναι κατάλληλη μιας και επιστρέφει συνεχώς πολύ χαμηλές τιμές. Ο βαθμός κάλυψης, επίσης, κυμαίνεται σε μέτρια επίπεδα, που σημαίνει ότι δεν πετυχαίνεται πάντα δημιουργία συστάσεων για κάθε πιθανό συνδυασμό.



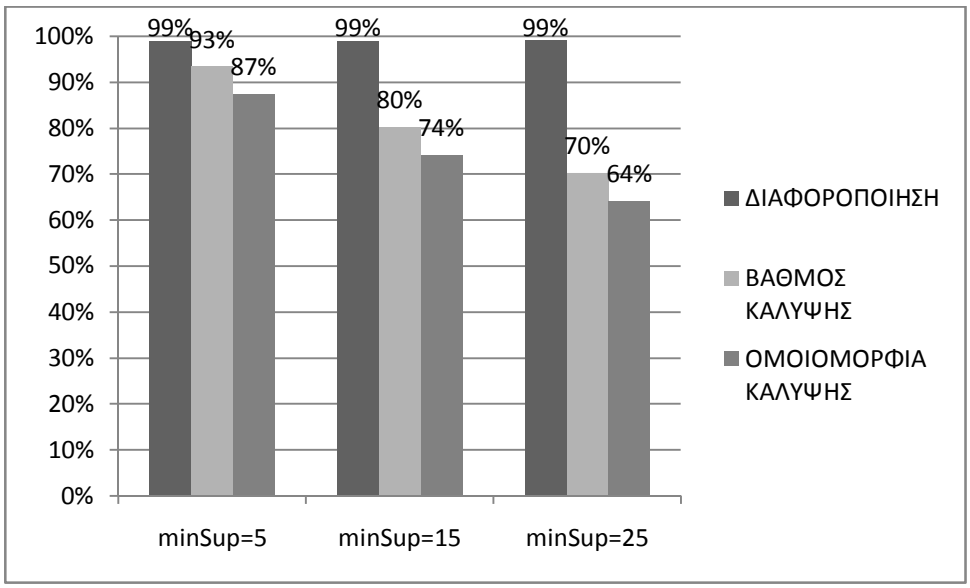
Σχήμα 4.9 Συστάσεις Βάσει Περιεχομένου (Καθολική Ανάλυση): Συσχετίσεις Διαφοροποίησης Και Κάλυψης Των Συστάσεων Λόγω Μεταβολών Της Υποστήριξης

Το συμπέρασμα που ανάγεται μέσα από τα πειραματικά αποτελέσματα για αυτή την προσέγγιση είναι ότι για μη συχνές εγγραφές στη βάση, εγγραφές με μικρή

επιλεκτικότητα, που παρόλα αυτά εμφανίζονται αρκετά συχνά ως συνδυασμοί μέσα στα αποτελέσματα της *Q*, και ανεξάρτητα της αυξομείωσης του βαθμού υποστήριξης, επιστρέφει αρκετά διαφοροποιημένες συστάσεις, χωρίς όμως να παρέχει το επιθυμητό πλήθος συστάσεων.

- *Συστάσεις Βάσει Σχήματος: Τοπική Ανάλυση*

Στις συστάσεις βάσει σχήματος, στην προσέγγιση της τοπικής ανάλυσης προχωρούμε στη δημιουργία διευρυμένων ερωτημάτων. Η αλλαγή θα επέλθει αρχικά με την επέκταση των γνωρισμάτων του `from`, και κατά επέκταση με προσθήκη κατάλληλων πεδίων στα γνωρίσματα του `select`, και κατάλληλων σχέσεων /κλειδιών των γνωρισμάτων του `where`.



Σχήμα 4.10 Συστάσεις Βάσει Σχήματος (Τοπική Ανάλυση): Συσχετίσεις Διαφοροποίησης Και Κάλυψης Των Συστάσεων Λόγω Μεταβολών Της Υποστήριξης

Τα αποτελέσματα του Σχήματος 4.10 προέρχονται από πειράματα που διεξήχθηκαν βάσει του μοτίβου των ερωτήσεων *Q*1 και *Q*2. Παρατηρούμε ότι για μικρές τιμές της υποστήριξης πετυχαίνουμε πολύ μεγάλα ποσοστά διαφοροποίησης συστάσεων, ομοιομορφίας κάλυψης αποτελεσμάτων, καθώς και βαθμού κάλυψης. Αυξάνοντας το βαθμό υποστήριξης, παρατηρούμε μία διαρκή πτώση τόσο του βαθμού κάλυψης όσο και της ομοιομορφίας κάλυψης των αποτελεσμάτων, αλλά παράλληλα σταθερότητα

στα ποσοστά της διαφοροποίησης. Επισημαίνουμε στο σημείο αυτό, ότι πλέον το B είναι αρκετά διευρυμένο σε σχέση με το A της ερώτησης του χρήστη και σε πολλές περιπτώσεις αυτή υπήρξε η κύρια αιτία της διαφοροποίησης των αποτελεσμάτων.

Όσον αφορά τα αποτελέσματα των πειραμάτων που βασίστηκαν στο μοτίβο των ερωτήσεων *Q*3 και *Q*4,, είχαμε τόσο κάλυψη όσο και διαφοροποίηση αποτελεσμάτων στο 100%, ανεξάρτητα από τη διακύμανση του βαθμού υποστήριξης. Εκείνο όμως που παρατηρήθηκε, είναι η διαφοροποίηση στην επέκταση των YMAL ερωτήσεων, σε σχέση με τις YMAL ερωτήσεις όπως διαμορφώθηκαν στην περίπτωση των συστάσεων βάσει σχήματος στην προσέγγιση της καθολικής ανάλυσης. Η επέκταση έγινε προς τελείως διαφορετικές κατευθύνσεις μέσα στους πίνακες της βάσης δεδομένων. Στην ανάλυση της επόμενης προσέγγισης, θα αναφέρουμε τους λόγους για τους οποίους παρατηρήθηκε αυτό το φαινόμενο.

- *Συστάσεις Βάσει Σχήματος: Καθολική Ανάλυση*

Όμοια με τις συστάσεις βάσει σχήματος στην τοπική ανάλυση, έτσι και στην προσέγγιση της καθολικής ανάλυσης, προχωρούμε στη δημιουργία διευρυμένων ερωτημάτων. Η αλλαγή επέρχεται ξανά με την επέκταση των γνωρισμάτων του `from`, και με τις επεκτάσεις των φράσεων του `select` και του `where`, που επηρεάζονται άμεσα. Εδώ αλλάζει όμως ο τρόπος με τον οποίο γίνεται η επέκταση στο S μέσω της συνιστώσας για την επιλογή σχέσεων.

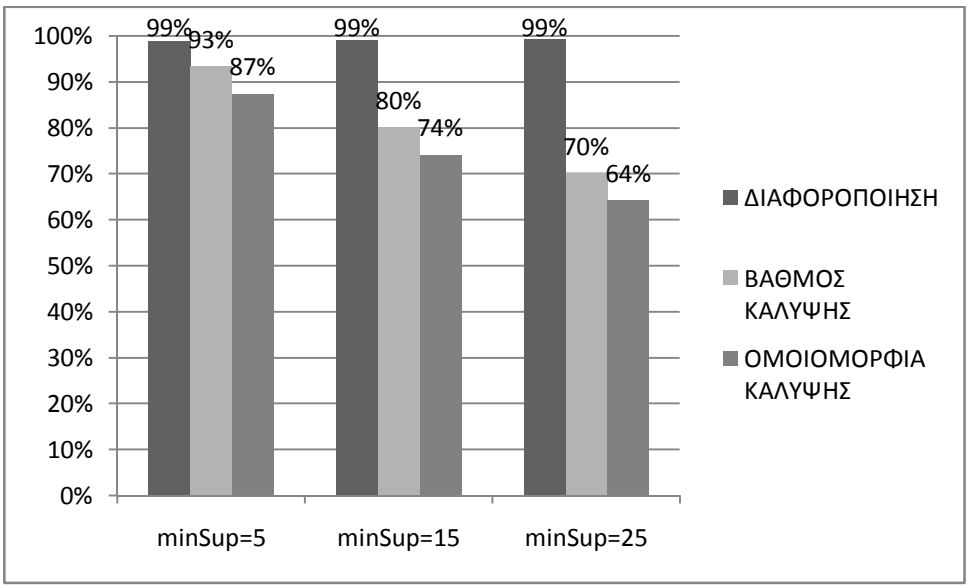
Πίνακας 4.3 Βαθμός *p* Μεταξύ Σχέσεων της *D*, για *k* =1

	Actor	Cast	Movies	Genre
Actor		0.21		
Cast	0.21		1.41	
Movies		1.41		1.62
Genre			1.62	
Movies-Cast	0.21			1.69

Όπως περιγράφτηκε αναλυτικά και πιο πάνω, για να υπολογίσουμε το S, διατηρούμε τις συσχετίσεις μεταξύ των σχέσεων στην *D*, οι οποίες χαρακτηρίζονται από έναν

βαθμό *p*. Πριν την επέκταση του S συμβουλευόμαστε αυτά τα δεδομένα (Πίνακας 4.3), και αν η σχέση R_j διατηρεί τον μέγιστο βαθμό *p*, τότε θα είναι τελικά εκείνη που θα επιλεγεί προς χρήση στα YMAL ερωτήματα, και η πληθικότητα των πινάκων στο S σε σχέση με το R θα αυξηθεί κατά ένα (*k* =1).

Τα αποτελέσματα του Σχήματος 4.11 προέρχονται από πειράματα που διεξήχθηκαν βάσει του μοτίβου των ερωτήσεων *Q*1 και *Q*2. Παρατηρούμε ότι τα αποτελέσματα είναι πανομοιότυπα με αυτά του Σχήματος 4.10. Αυτό οφείλεται και εδώ στο γεγονός ότι το B είναι αρκετά διευρυμένο σε σχέση με το A της ερώτησης του χρήστη. Το διευρυμένο B υπονοεί πλειάδες που ενδέχεται να εμφανίζουν επαναλαμβανόμενο ένα ή περισσότερα χαρακτηριστικά, αλλά όμως στο σύνολό τους διαφοροποιούνται μεταξύ τους.



Σχήμα 4.11 Συστάσεις Βάσει Σχήματος (Καθολική Ανάλυση): Συσχετίσεις Διαφοροποίησης Και Κάλυψης Των Συστάσεων Λόγω Μεταβολών Της Υποστήριξης

Τα αποτελέσματα των πειραμάτων που βασίστηκαν στο μοτίβο των ερωτήσεων *Q*3 και *Q*4, παρουσίασαν και σε αυτή την προσέγγιση κάλυψη και διαφοροποίηση με ποσοστό 100%, ανεξάρτητα από τη διακύμανση του βαθμού υποστήριξης. Όμως, η επέκταση του S έγινε προς τελείως διαφορετικές κατευθύνσεις μέσα στους πίνακες της βάσης δεδομένων, σε σχέση με την προηγούμενη προσέγγιση. Αυτό οφείλεται

στον διαφορετικό τρόπο με τον οποίο επεκτείνουμε το S. Βάσει της κατανομής των εγγραφών στη βάση δεδομένων, είναι εύκολο να διαπιστωθεί ότι η κατεύθυνση της επέκταση του S στην καθολική προσέγγιση είναι κάτι δεδομένο με βάση την τρέχουσα ερώτηση *Q*. Αντίθετα, στην τοπική ανάλυση, η επέκταση γίνεται δυναμικά κάθε φορά με βάση τα ίδια τα αποτελέσματα της *Q* και το βαθμό των συνενώσεων που πετυχαίνουν με τους γειτονικούς τους πίνακες. Σε αυτό το λόγο έγκειται η διαφοροποίηση των διευρυμένων ερωτημάτων που μελετήσαμε στα πειράματά μας.

Το συμπέρασμα που ανάγεται μέσα από τα πειραματικά αποτελέσματα συνολικά για τις συστάσεις βάσει σχήματος και για εγγραφές με μεγάλη επιλεκτικότητα, είναι ότι για μικρές τιμές του βαθμού υποστήριξης, επιστρέφει αρκετά 'χρήσιμες' συστάσεις οι οποίες είναι διαφοροποιημένες μεταξύ τους, και καλύπτουν σχεδόν κάθε πιθανό συνδυασμό, με $M' \subseteq M$, και $m_i \subseteq m$. Η ανάλυση ευαισθησίας μας υποδεικνύει ως καταλληλότερο βαθμό υποστήριξης το *5*.

4.2.3. Περιπτώσεις παρόμοιων συστάσεων ανάμεσα στις διαφορετικές προσεγγίσεις

Η φύση των συστάσεων βάσει περιεχομένου μας οδήγησε στο συμπέρασμα ότι είναι πιθανό να υπάρξουν συστάσεις με το ίδιο περιεχόμενο σε πολύ μεγάλο ποσοστό, είτε από την προσέγγιση της τοπικής ανάλυσης, είτε από την προσέγγιση της καθολικής ανάλυσης. Τέτοιου είδους όμοιες συστάσεις είναι δυνατόν να υπάρξουν όταν στην *Q* περιέχονται μόνο τιμές χαρακτηριστικών που είναι πολύ σπάνιες στη βάση δεδομένων, και ο βαθμός υποστήριξης είναι αρκετά μικρός ώστε να υπάρξει κάλυψη αποτελεσμάτων. Πράγματι, μέσα από μία σειρά πειραμάτων καταφέραμε να αποδείξουμε αυτή την υπόθεση.

Έστω, λοιπόν, η ερώτηση του χρήστη *Q*4:

```
select  GENRE.genre
from    CAST, ACTOR, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        CAST.pid=ACTOR.pid and
        ACTOR.lname = 'Jolie';
```

Βάσει της *Q*4 δημιουργήσαμε ομοίως άλλες *10* παρόμοιες ερωτήσεις αλλάζοντας το επίθετο του ηθοποιού κάθε φορά, επιλέγοντας μόνο εκείνους που δεν εμφανίζονται πολύ συχνά στη βάση δεδομένων. Λόγω του ότι η βάση δεδομένων που χρησιμοποιούμε για την εφαρμογή της μεθόδου των YMAL συστάσεων περιέχει αρκετούς ηθοποιούς του παλιού κλασικού αμερικάνικου κινηματογράφου, τα ονόματα των πιο σύγχρονων ηθοποιών αποτελούν σπάνιες εγγραφές στη βάση δεδομένων. Για όλα τα πειράματα χρησιμοποιήσαμε τον βαθμό υποστήριξης ίσο με *2*. Στον Πίνακα 4.4 διακρίνονται τα ποσοστά ομοιότητας των αποτελεσμάτων στις δύο προσεγγίσεις.

Πίνακας 4.4 Ομοιότητα YMAL Συστάσεων Βάσει Περιεχομένου Για Σπάνιες Εγγραφές Στη Βάση Δεδομένων

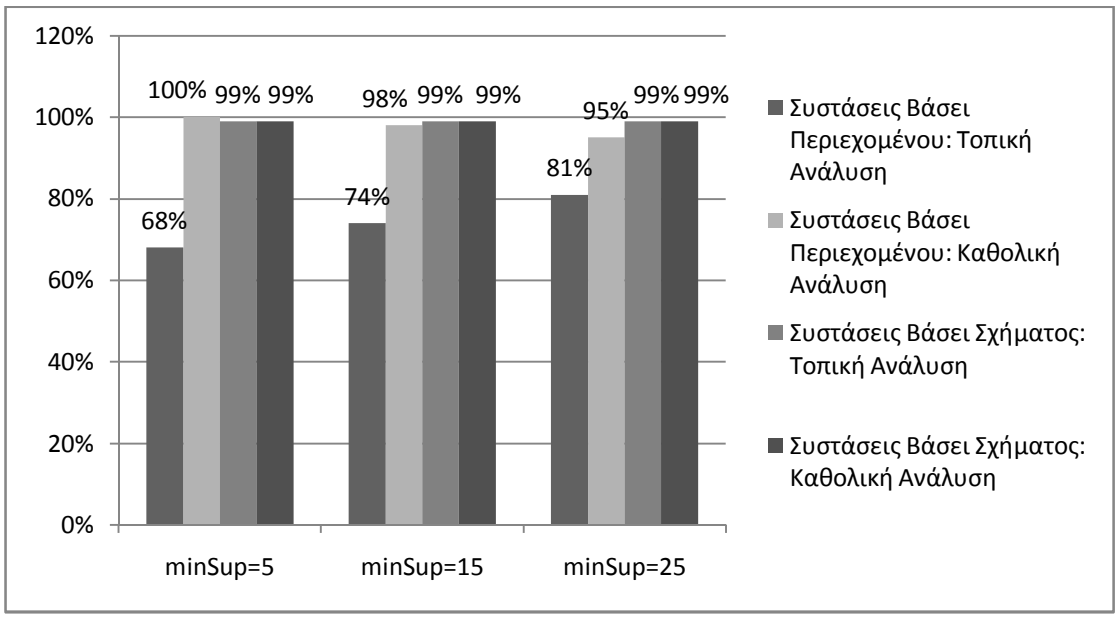
	Ομοιότητα Συστάσεων Μεταξύ των Δύο Προσεγγίσεων	Συχνότητα Εμφάνισης Εγγραφής Στη Βάση Δεδομένων	Σύνολο Αποτελεσμάτων Στο $R(Q)$
Jolie	90%	2	58
Pitt	90%	24	213
Paltrow	90%	1	74
Cage	90%	5	108
Cruse	90%	7	11
Kidman	90%	1	77
Depp	80%	3	211
Carrey	80%	6	78
Banderas	70%	5	106
Stiller	70%	10	170

Για τις συστάσεις βάσει σχήματος, όπως είδαμε και στην προηγούμενη υποενότητα, είναι δύσκολο να βγάλουμε ένα γενικό συμπέρασμα για το πότε τα αποτελέσματα των συστάσεων θα είναι σε μεγάλο βαθμό όμοια. Αυτό εξαρτάται τόσο από το βαθμό *p* των σχέσεων μέσα στην ίδια τη *D*, όσο και με το βαθμό *p* που προκύπτει δυναμικά από την συνένωση των υποσυνόλων της βάσης δεδομένων, όπως προκύπτουν μέσα από το $R(Q)$, σε σχέση με τις ίδιες τις σχέσεις της *D*.

4.3. Σύγκριση Αποτελεσμάτων Διαφορετικών Προσεγγίσεων

Σε αυτή την ενότητα, θα παρουσιάσουμε συνολικά τα αποτελέσματα των πειραμάτων μας, ομαδοποιώντας τα σε σχέση με τις μετρικές που μελετάμε και σχολιάζοντας τις επιδόσεις της κάθε προσέγγισης.

Το Σχήμα 4.13 αναφέρεται στη διαφοροποίηση YMAL συστάσεων κατά την αύξηση του βαθμού υποστήριξης. Εδώ παρατηρούμε ότι οι προσεγγίσεις των συστάσεων βάσει σχήματος, αλλά και η καθολική ανάλυση στην προσέγγιση των συστάσεων βάσει περιεχομένου πετυχαίνουν πολύ ικανοποιητικά ποσοστά διαφοροποίησης, τα οποία διατηρούν ανεξάρτητα με την αύξηση του βαθμού υποστήριξης. Θυμίζουμε ότι η ομάδα των συγκεκριμένων πειραμάτων αναφέρεται σε δημοφιλείς εγγραφές μέσα στη βάση δεδομένων.

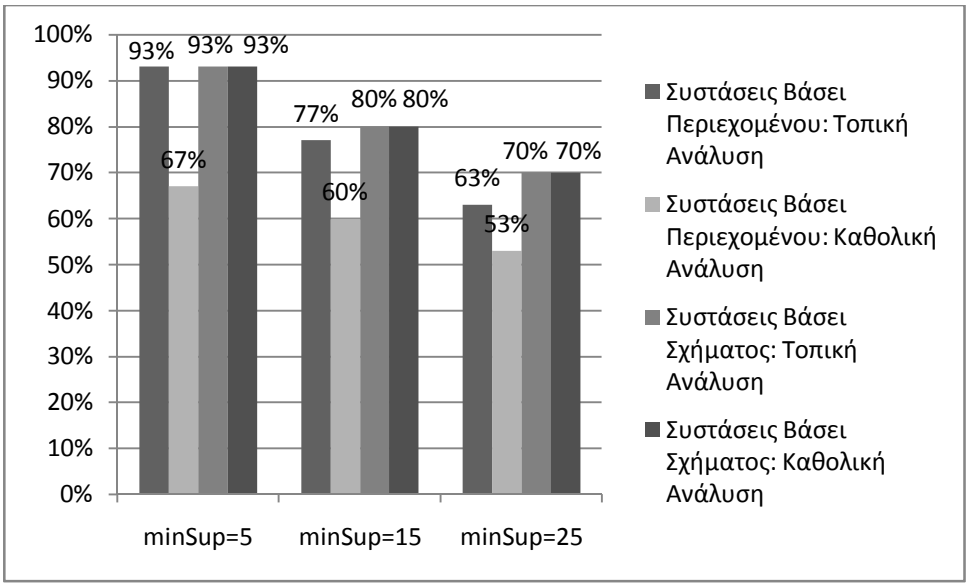


Σχήμα 4.12 Διαφοροποίηση YMAL Συστάσεων

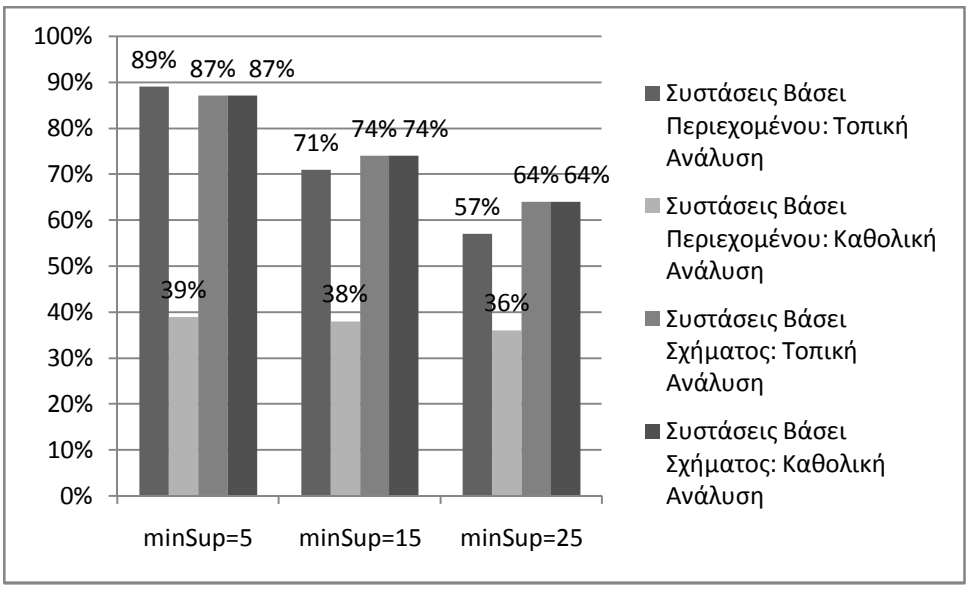
Στο Σχήμα 4.14 παρατηρούνται τα ποσοστά του βαθμού κάλυψης των YMAL συστάσεων κατά την αύξηση της υποστήριξης. Εδώ φαίνεται ότι τα καλύτερα ποσοστά λαμβάνονται στις προσεγγίσεις των συστάσεων βάσει σχήματος, αλλά και στην τοπική ανάλυση στην προσέγγιση των συστάσεων βάσει περιεχομένου, μόνο όμως για τις περιπτώσεις εκείνες που ο βαθμός υποστήριξης είναι σχετικά μικρός.

Στην αύξηση του βαθμού υποστήριξης, τα ποσοστά του βαθμού κάλυψης των συστάσεων μειώνονται σημαντικά.

Από την άλλη μεριά, στο Σχήμα 4.15 παρατηρούνται τα ποσοστά ομοιομορφίας της κάλυψης των YMAL συστάσεων. Τα αποτελέσματα συμβαδίζουν με εκείνα του Σχήματος 4.14. Για τις προσεγγίσεις των συστάσεων βάσει σχήματος, αλλά και στην τοπική ανάλυση στην προσέγγιση των συστάσεων βάσει περιεχομένου, και μόνο για τις περιπτώσεις εκείνες που ο βαθμός υποστήριξης είναι σχετικά μικρός, πετυχαίνουμε ικανοποιητικά ποσοστά ομοιομορφίας της κάλυψης των συστάσεων. Στην αύξηση του βαθμού υποστήριξης, τα ποσοστά ομοιομορφίας της κάλυψης των συστάσεων και πάλι μειώνονται σημαντικά.



Σχήμα 4.13 Βαθμός Κάλυψης YMAL Συστάσεων



Σχήμα 4.14 Ομοιομορφία Κάλυψης YMAL Συστάσεων

Μέσω των παρατηρήσεων που προέρχονται από το σύνολο των πειραματικών αποτελεσμάτων, καταλήγουμε στα ακόλουθα χρήσιμα συμπεράσματα. Για μεγάλα συστήματα βάσεων δεδομένων όπου οι όμοιες εγγραφές στους πίνακες είναι πολύ σπάνιες θα ήταν καλό να εφαρμοστούν προσεγγίσεις συστάσεων βάσει σχήματος, τόσο η τοπική ανάλυση όσο και η καθολική, ή ακόμη και ένα συνδυασμός τους δημιουργώντας έτσι ένα νέο υβριδικό μοντέλο, με χρήση μικρών βαθμών υποστήριξης. Για συστήματα δεδομένων με αρκετές δημοφιλείς εγγραφές, προτείνουμε επιπλέον την εφαρμογή της προσέγγισης των συστάσεων βάσει περιεχομένου με τοπική ανάλυση των δεδομένων και ενδιάμεσες τιμές σε βαθμούς υποστήριξης. Τέλος, για συστήματα δεδομένων με τις πιο σπάνιες εγγραφές να υπερισχύουν σε σχέση με τις δημοφιλείς, η εφαρμογή της προσέγγισης των συστάσεων βάσει περιεχομένου τόσο με χρήση τοπικής ανάλυσης των δεδομένων όσο και καθολικής ανάλυσης, θα επέφερε αρκετά όμοιο σύνολο συστάσεων.

ΚΕΦΑΛΑΙΟ 5. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Στο τελευταίο κεφάλαιο αυτής της διατριβής, κάνουμε μία ανακεφαλαίωση για το σύνολο της εργασίας μας, και μέσω αυτής παρουσιάζουμε τους στόχους που θέσαμε αρχικά και πώς αυτοί υλοποιήθηκαν, συνάγοντας χρήσιμα συμπεράσματα. Επιπλέον, παρουσιάζουμε διάφορες προτάσεις για την επέκταση της εφαρμογή της μεθόδου που χρήζουν περαιτέρω έρευνας.

5.1. Συμπεράσματα

Στην εργασία αυτή μελετήσαμε διάφορες προσεγγίσεις για τη δημιουργία συστάσεων σε σχεσιακές βάσεις δεδομένων, θεωρώντας ως συστάσεις προς τους χρήστες πλειάδες που δεν ανήκουν στο αποτέλεσμα της ερώτησής τους, αλλά που πιθανότατα είναι ενδιαφέρουσες για αυτούς. Μέσω της τεχνικής δημιουργίας συστάσεων βασισμένων στην τρέχουσα κατάσταση των δεδομένων, διακρίναμε δύο μεγάλες κατηγορίες: τις συστάσεις βάσει περιεχομένου και τις συστάσεις βάσει σχήματος. Κάθε μία από αυτές τις δύο κατηγορίες μπορεί να προσεγγιστεί είτε μέσω τοπικής ανάλυσης, είτε μέσω καθολικής ανάλυσης των δεδομένων που παρέχονται.

Εν συνεχεία, ορίσαμε το μοντέλο των YMAL συστάσεων για κάθε μία κατηγορία, με τις υποκατηγορίες της. Προτείναμε μία κοινή αρχιτεκτονική για κάθε κατηγορία, μέσω της οποίας διασφαλίζεται η έγκυρη επεξεργασία των ερωτήσεων των χρηστών και η απόρροια συστάσεων. Καθορίσαμε τις μεταβλητές που τις επηρεάζουν και παρουσιάσαμε τους αναγκαίους αλγορίθμους που χρειάστηκαν για τον υπολογισμό

των αποτελεσμάτων. Βάσει αυτών, δημιουργήσαμε ένα ολοκληρωμένο πλαίσιο, το οποίο παρέχει τη δυνατότητα εφαρμογής κάθε μίας περίπτωσης ξεχωριστά. Πειραματιζόμενοι πάνω σε αυτές, παρατηρήσαμε ότι παρουσιάζουν διαφορές σχετικά με τα επίπεδα κάλυψης και διαφοροποίησης των αποτελεσμάτων που επιστρέφουν ως συστάσεις προς τους χρήστες του συστήματος, εντοπίζοντας τις κύριες αιτίες κάθε φορά.

Σαν τελικό συμπέρασμα, μπορούμε να πούμε ότι οι στόχοι που θέσαμε στην αρχή αυτής της εργασίας ικανοποιήθηκαν σε πολύ μεγάλο βαθμό. Οι YMAL συστάσεις είναι στο σύνολό τους πλήρεις, πετυχαίνοντας κάλυψη και διαφοροποίηση αποτελεσμάτων, σύμφωνα με τις αρχικές απαιτήσεις μας, ενώ οποιαδήποτε άλλη πληροφορία παρουσιάζεται μέσα από το ολοκληρωμένο πλαίσιο, θα ήταν πολύτιμη και σε πραγματικά σενάρια χρήσης.

5.2. Μελλοντική Εργασία

Οι πειραματικές μετρήσεις του YMAL συστήματος έγιναν με βάση τη μέτρηση του βαθμού κάλυψης, της ομοιομορφίας κάλυψης και της διαφοροποίησης των συστάσεων. Θα ήταν πολύ ενδιαφέρουσα η δοκιμή του συστήματος σε πραγματικούς χρήστες, οι οποίοι μέσω κατάλληλης φόρμας θα μπορούσαν να δίνουν ένα βαθμό αρεσκείας στις YMAL συστάσεις. Αυτός ο βαθμός αρεσκείας θα μπορούσε να χρησιμοποιηθεί ανεξάρτητα από το σύστημα, για να υποδείξει ποια είναι η καλύτερη προσέγγιση συστάσεων με βάση τη γνώμη της πλειοψηφίας των χρηστών. Επιπλέον, μία άλλη σκέψη είναι να μπορούσε αυτός ο βαθμός αρεσκείας των χρηστών να ενσωματωθεί στην υπάρχουσα αρχιτεκτονική του συστήματος, και να λαμβανόταν υπόψη κατά τη σύσταση των YMAL αποτελεσμάτων.

Επιπρόσθετα, ενδιαφέρον παρουσιάζει η ιδέα ανάπτυξης υβριδικής εφαρμογής για την κατασκευή των YMAL συστάσεων, τόσο για την προσέγγιση βάσει περιεχομένου, όσο και για την προσέγγιση βάσει σχήματος, όπου θα συνδυάζονταν η τοπική και καθολική ανάλυση. Η αρχιτεκτονική του YMAL συστήματος συστάσεων, μας οδήγησε στην κατασκευή ενός πλαισίου εφαρμογών ανοιχτού κώδικα και εύκολα μετατρέψιμου.

ΑΝΑΦΟΡΕΣ

[1] Adomavicius, G. and Tuzhilin, A., "Multidimensional Recommender Systems: A Data Warehousing Approach", In Proceedings of the International Workshop on Electronic Commerce, pp. 180-192, 2001

[2] Gediminas Adomavicius, Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 6, June 2005.

[3] Kostas Stefanidis, Marina Drosou, and Evaggelia Pitoura, "You May Also Like: Results in Relational Databases", ACM VLDB 2009.

[4] Vagelis Hristidis, Yannis Papakonstantinou, "DISCOVER: Keyword Search in Relational Databases", Proceeding of the 28th VLDB Conference, Hong Kong, China, 2002.

[5] Kostas Stefanidis, Marina Drosou and Evaggelia Pitoura, "PerK: Personalized Keyword Search in Relational Databases through Preferences", EDBT 2010, March 22-26, 2010, Lausanne, Switzerland.

[6] Gloria Chatzopoulou, Magdalini Eirinaki, and Neoklis Polyzotis, "Query Recommendations for Interactive Database Exploration". In SSDBM 2009, pp. 3-18, 2009.

[7] Burke, R., "Hybrid Recommender Systems: Survey and Experiments", Journal of User Modeling and User-Adapted Interaction, Vol. 12, pp. 331-370, 2002.

[8] Koutrika, G., Bercovitz, B. and Garcia-Molina, H., "FlexRecs: Expressing and Combining Flexible Recommendations", In Proceedings of ACM SIGMOD Conference, pp. 745-757, 2009.

[9] Chaudhuri S., Das G., Hristidis V., Weikum G., "Probabilistic Information Retrieval Approach for Ranking of Database Query Results", ACM Transactions on Database Systems, Vol. 31, No.3, pp.1134-1168, September 2006.

[10] Agrawal R, Srikant R, 'Fast Algorithms for mining association rules'. In Proceedings of the 1994 international conference on very large data bases (VLDB'94), Santiago, Chile, pp 487–499.

[11] Park JS, Chen MS, Yu PS, “Efficient parallel mining for association rules”, In Proceeding of the 4th international conference on information and knowledge management, Baltimore, MD, pp 31–36, 1995.

[12] Savasere A, Omiecinski E, Navathe S, “An efficient algorithm for mining association rules in large databases”, In Proceeding of the 1995 international conference on very large data bases (VLDB’95), Zurich, Switzerland, pp 432–443, 1995.

[13] C. Yu, L. V. S. Lakshmanan, S. Amer-Yahia, “Recommending diversification using explanations”, In ICDE, pp 1299-1302, 2009.

[14] M. Trick, “Sensitivity Analysis For Linear Programming”, Available at http://mat.gsia.cmu.edu/classes/quant/notes/chap8/chap8.html

[15] Wikipedia, “The free encyclopedia”, Available at http://en.wikipedia.org.

[16] LibraryThing, “Catalogs Your Books Online”, Available at www.librarything.com.

[17] IMDB, “Internet Movies Database”, Available at www.imdb.com.

ΠΑΡΑΡΤΗΜΑ

Σε αυτό το σημείο παραθέρουμε ενδεικτικά αποτελέσματα από τα πειράματα που χρησιμοποιήσαμε σε αυτή τη διατριβή καθώς και ορισμένες από τις SQL ερωτήσεις στις οποίες βασιστήκαμε.

Ερωτήσεις Βασισμένες στο Μοτίβο της *Q*1:

```
select  CAST.role, GENRE.genre
from    ACTOR, CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        ACTOR.pid = CAST.pid and
        ACTOR.fname= 'Lee' and
        ACTOR.lname= 'Phelps';
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%
minSup=5	30	70%	100%	100%
minSup=15	20	90%	67%	67%
minSup=25	20	90%	67%	67%
GLOBAL				
ΙΔΑΝΙΚΑ	30	100%	100%	100%
minSup=5	18	100%	67%	60%
minSup=15	18	94%	67%	60%
minSup=25	9	100%	33%	30%

SCHEMA LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%
minSup=5	30	100%	100%	100%
minSup=15	20	100%	67%	67%
minSup=25	20	100%	67%	67%
SCHEMA GLOBAL				
ΙΔΑΝΙΚΑ	30	100%	100%	100%
minSup=5	30	100%	100%	100%
minSup=15	20	100%	67%	67%
minSup=25	20	100%	67%	67%

```
select  CAST.role, GENRE.genre
from    ACTOR, CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        ACTOR.pid = CAST.pid and
        ACTOR.fname= 'Mel' and
        ACTOR.lname= 'Blanc';
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	SCHEMA LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	12	92%	100%	40%	**minSup=5**	12	92%	100%	40%
minSup=15	12	92%	100%	40%	**minSup=15**	12	92%	100%	40%
minSup=25	12	92%	100%	40%	**minSup=25**	12	92%	100%	40%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	12	92%	100%	40%
minSup=15	9	100%	67%	30%	**minSup=15**	12	92%	100%	40%
minSup=25	9	100%	67%	30%	**minSup=25**	12	92%	100%	40%

```
select  CAST.role, GENRE.genre
from    ACTOR, CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        ACTOR.pid = CAST.pid and
        ACTOR.fname= 'Bess' and
        ACTOR.lname= 'Flowers';
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	SCHEMA LOCAL	ΣΥΝ-ΟΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	30	87%	100%	100%	**minSup=5**	30	97%	100%	100%
minSup=15	30	87%	100%	100%	**minSup=15**	30	97%	100%	100%
minSup=25	20	100%	67%	67%	**minSup=25**	20	100%	67%	67%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	10	100%	67%	33%	**minSup=5**	30	97%	100%	100%
minSup=15	18	100%	67%	60%	**minSup=15**	30	97%	100%	100%
minSup=25	9	100%	33%	30%	**minSup=25**	20	100%	67%	67%

```
select  CAST.role, GENRE.genre
from    ACTOR, CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        ACTOR.pid = CAST.pid and
        ACTOR.fname= 'James' and
        ACTOR.lname= 'Flavin';
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	SCHEMA LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ

ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	30	73%	100%	100%	**minSup=5**	30	100%	100%	100%
minSup=15	20	90%	67%	67%	**minSup=15**	20	100%	67%	67%
minSup=25	10	100%	33%	33%	**minSup=25**	10	100%	33%	33%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	30	100%	100%	100%
minSup=15	9	100%	33%	30%	**minSup=15**	20	100%	67%	67%
minSup=25	9	100%	33%	30%	**minSup=25**	10	100%	33%	33%

```
select  CAST.role, GENRE.genre
from    ACTOR, CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        ACTOR.pid = CAST.pid and
        ACTOR.fname= 'Irving' and
        ACTOR.lname= 'Bacon';
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	**SCHEMA LOCAL**	ΣΥΝΟΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	30	90%	100%	100%	**minSup=5**	30	100%	100%	100%
minSup=15	10	100%	33%	33%	**minSup=15**	30	100%	33%	33%
minSup=25	10	100%	33%	33%	**minSup=25**	10	100%	33%	33%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	30	100%	100%	100%
minSup=15	9	100%	33%	30%	**minSup=15**	30	100%	33%	33%
minSup=25	9	100%	33%	30%	**minSup=25**	10	100%	33%	33%

Ερωτήσεις Βασισμένες στο Μοτίβο της *Q*2:

```
select  CAST.role, MOVIES.year
from    CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        GENRE.genre= 'Western'
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	**SCHEMA LOCAL**	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%

minSup=5	30	47%	100%	100%	**minSup=5**	30	100%	100%	100%
minSup=15	30	47%	100%	100%	**minSup=15**	30	100%	100%	100%
minSup=25	30	47%	100%	100%	**minSup=25**	30	100%	100%	100%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	30	100%	100%	100%
minSup=15	9	100%	67%	30%	**minSup=15**	30	100%	100%	100%
minSup=25	9	100%	67%	30%	**minSup=25**	30	100%	100%	100%

```
select  CAST.role, MOVIES.year
from    CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        GENRE.genre= 'Fantasy'
```

LOCAL	ΣΥΝΟΛΟ	ΔΙΑΦΟΡΟΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	**SCHEMA LOCAL**	ΣΥΝΟΛΟ	ΔΙΑΦΟΡΟΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	20	60%	67%	67%	**minSup=5**	20	100%	67%	67%
minSup=15	20	60%	67%	67%	**minSup=15**	20	100%	67%	67%
minSup=25	20	60%	67%	67%	**minSup=25**	20	100%	67%	67%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	20	100%	67%	67%
minSup=15	9	100%	67%	30%	**minSup=15**	20	100%	67%	67%
minSup=25	9	100%	67%	30%	**minSup=25**	20	100%	67%	67%

```
select  CAST.role, MOVIES.year
from    CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        GENRE.genre= 'Mystery'
```

LOCAL	ΣΥΝΟΛΟ	ΔΙΑΦΟΡΟΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	**SCHEMA LOCAL**	ΣΥΝΟΛΟ	ΔΙΑΦΟΡΟΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	30	47%	100%	100%	**minSup=5**	30	100%	100%	100%
minSup=15	20	65%	67%	67%	**minSup=15**	20	100%	100%	100%
minSup=25	20	65%	67%	67%	**minSup=25**	20	100%	100%	100%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚ**	30	100%	100%	100%

					A				
minSup=5	9	100%	67%	30%	**minSup=5**	30	100%	100%	100%
minSup=15	15	87%	67%	50%	**minSup=15**	20	100%	100%	100%
minSup=25	18	83%	67%	60%	**minSup=25**	20	100%	100%	100%

```
select  CAST.role, MOVIES.year
from    CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        GENRE.genre= 'Adventure'
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	**SCHEMA LOCAL**	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	20	60%	67%	67%	**minSup=5**	20	100%	67%	67%
minSup=15	20	60%	67%	67%	**minSup=15**	20	100%	67%	67%
minSup=25	20	60%	67%	67%	**minSup=25**	20	100%	67%	67%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	20	100%	67%	67%
minSup=15	9	100%	67%	30%	**minSup=15**	20	100%	67%	67%
minSup=25	18	67%	67%	60%	**minSup=25**	20	100%	67%	67%

```
select  CAST.role, MOVIES.year
from    CAST, GENRE, MOVIES
where   GENRE.mid = MOVIES.mid and
        CAST.mid=MOVIES.mid and
        GENRE.genre= 'Musical'
```

LOCAL	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ	**SCHEMA LOCAL**	ΣΥΝΟ-ΛΟ	ΔΙΑΦΟΡΟ-ΠΟΙΗΣΗ	ΒΑΘΜΟΣ ΚΑΛΥΨΗΣ	ΟΜΟΙΟ-ΜΟΡΦΙΑ ΚΑΛΥΨΗΣ
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	30	53%	100%	100%	**minSup=5**	30	100%	100%	100%
minSup=15	30	53%	100%	100%	**minSup=15**	30	100%	100%	100%
minSup=25	10	100%	33%	33%	**minSup=25**	20	100%	67%	67%
GLOBAL					**SCHEMA GLOBAL**				
ΙΔΑΝΙΚΑ	30	100%	100%	100%	**ΙΔΑΝΙΚΑ**	30	100%	100%	100%
minSup=5	9	100%	67%	30%	**minSup=5**	30	100%	100%	100%
minSup=15	9	100%	67%	30%	**minSup=15**	30	100%	100%	100%
minSup=25	9	100%	67%	30%	**minSup=25**	20	100%	67%	67%

Ερωτήσεις Βασισμένες στο Μοτίβο της *Q*3:

```
select  CAST.role
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        MOVIES.year = '2003';
```

```
select  CAST.role
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        MOVIES.year = '2000';
```

```
select  CAST.role
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        MOVIES.year = '1956';
```

```
select  CAST.role
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        MOVIES.year = '1937';
```

```
select  CAST.role
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        MOVIES.year = '1945';
```

Ερωτήσεις Βασισμένες στο Μοτίβο της *Q*4:

```
select  MOVIES.year
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        CAST.role = 'Policeman';
```

```
select  MOVIES.year
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        CAST.role = 'Detective';
```

```
select  MOVIES.year
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        CAST.role = 'Reporter';
```

```
select  MOVIES.year
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        CAST.role = 'Secretary';
```

```
select  MOVIES.year
from    CAST, MOVIES
where   CAST.mid=MOVIES.mid and
        CAST.role = 'Journalist';
```

ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΥΓΓΡΑΦΕΑ

[1] Eftychia Koletsou, Kostas Stefanidis, Marina Drosou, Evaggelia Pitoura, “YMAL: Ένα Σύστημα Συστάσεων Σχεσιακών Βάσεων Δεδομένων”, HDMS 2010.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Η Ευτυχία Κωλέτσου γενήθηκε στην Άρτα το 1983. Το 2005 αποφοίτσε από το Τμήμα Ψηφιακών Συστημάτων του Πανεπιστημιού Πειραιώς. Κατά τη χρονική περίοδο 08/2007-02/2008 εργάστηκε ως μέλος της ερευνητικής ομάδας OSIRIS στο TIMC-IMAG στη Grenoble της Γαλλίας. Το ακαδημαϊκό έτος 2008-2009 έγινε δεκτή στο Πρόγραμμα Μεταπτυχιακών Σπουδών του Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων, ενώ από τον Σεπτέμβριο του 2009 είναι μέλος του Distributed Management of Data Laboratory (DMOD) του ιδίου τμήματος. Κατέχει θέση Εργαστηριακού Συνεργάτη στο Ανώτατο Τεχνολογικό Ίδρυμα Ηπείρου, στο Τμήμα Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών και στο Τμήμα Λογιστικής.

Τα ερευνητικά της ενδιαφέροντα κινούνται στους χώρους των βάσεων δεδομένων, των συστημάτων σύστασης, της ανάκτησης πληροφορίας, και της ασφάλειας των πληροφοριακών συστημάτων.