ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη Πληροφορίας από τα Δεδομένα Χρήσης του HEAL-Link portal

Του φοιτητή

Νικόλαου Παράσχου

Αρ. Μητρώου: 03/2245

Επιβλέπων καθηγητής

Δημήτρης Α. Δέρβος

Θεσσαλονίκη 2010

Περίληψη

Ο διαδικτυακός τόπος HEAL-Link (http://www.heal-link.gr/journals/) συνιστά κομβικό σημείο διαδικτυακής πρόσβασης της Ελληνικής και της Κυπριακής ακαδημαϊκής κοινότητας σε δεκάδες χιλιάδες τίτλους επιστημονικών περιοδικών, σε ηλεκτρονική/ψηφιακή μορφή. Λειτουργώντας επί μία δεκαετία περίπου, η βάση δεδομένων του HEAL-Link διαθέτει σήμερα εξαιρετικά πλούσια δεδομένα χρήσης (log files) του περιβάλλοντος. Στην πορεία, η διαδικτυακή εφαρμογή της βάσης δεδομένων έχει εξελιχθεί/βελτιωθεί και ως προς τον κώδικά της και ως προς την έκδοση του συστήματος διαχείρισής της (IBM DB2). Σήμερα, τα δεδομένα του HEAL-Link υπόκεινται στο πλέον σύγχρονο είδος διαχείρισης και επεξεργασίας διαδικτυακού συστήματος το οποίο συμπεριλαμβάνει διακομιστή web (web server), διακομιστή εφαρμογής (application server) και διαχειριστή βάσεων δεδομένων (DBMS server).

Στόχο της συγκεκριμένης πτυχιακής εργασίας συνιστά η επεξεργασία των δεδομένων χρήσης του HEAL-Link προς εξόρυξη πληροφορίας η σημασία/αξία της οποίας εξυπηρετεί δύο ανάγκες: α) τη διαδικασία λήψης αποφάσεων από το γραφείο διαχείρισης του διαδικτυακού τόπου HEAL-Link, και β) την εξυπηρέτηση των χρηστών του διαδικτυακού τόπου στην αναζήτηση επιστημονικής βιβλιογραφίας σχετικής προς το/α γνωστικό αντικείμενο/α της εξειδίκευσής τους. Ειδικά για την (β) ανάγκη, μέρος της πτυχιακής εργασίας συνιστά η ανάπτυξη πιλοτικού συστήματος αυτόματης παραγωγής συστάσεων (recommender system) όπου, π.χ., ανάλογα με τα περιοδικά ή/και θέματα που στοχοποιεί ένας χρήστης, το σύστημα του προτείνει και άλλα τα οποία άλλοι χρήστες έχουν στοχοποιήσει σε ανάλογου τύπου αναζητήσεις.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στον καθηγητή, κ.Δ.Δέρβο, για την πολύτιμη καθοδήγησή του στην εκπόνηση της παρούσης εργασίας. Επίσης, είμαι ιδιαίτερα ευγνώμων στους διαχειριστές της διαδικτυακής πύλης του HEAL-Link, Πόπη Αξονίδου και Λεωνίδα Πισπιρίγγα, για τις συμβουλές και τις απαντήσεις τους στα ερωτήματα που έθετα. Τέλος και πλέον σημαντικώς, θα ήθελα να ευχαριστήσω τους γονείς μου για την αμέριστη υποστήριξη και υπομονή τους καθόλη τη διάρκεια εκπόνησης της εργασίας.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος

Στην εργασία αυτή εφαρμόζουμε τη διεργασία εξόρυξης πληροφορίας στα δεδομένα χρήσης του HEAL-Link portal. Χρησιμοποιούμε λογισμικό της IBM για να εξορύξουμε χρήσιμες πληροφορίες από πραγματικά δεδομένα. Εφαρμόζουμε βήμα βήμα όλες τις φάσεις της διεργασίας εξόρυξης, περιγράφοντας αναλυτικά τις εκτελούμενες ενέργειες και τον τρόπο χρήσης του λογισμικού.

Ξεκινάμε με την κατανόηση της επιχείρησης και των επιχειρηματικών προβλημάτων και συνεχίζουμε με την μετατροπή των τελευταίων σε προβλήματα εξόρυξης πληροφορίας, τα οποία και επιλύουμε. Αφού πρώτα εξερευνήσουμε τα δεδομένα και τα προετοιμάσουμε κατάλληλα προς εξόρυξη, τα τροφοδοτούμε στη συνέχεια στον αλγόριθμο εξόρυξης ο οποίος παράγει τα ζητούμενα μοντέλα. Αξιολογούμε τα παραγόμενα αποτελέσματα και παρουσιάζουμε τον τρόπο ανάπτυξης μίας λειτουργικής λύσης για τη μόνιμη διάθεσή τους στην εφαρμογή του συστήματος HEAL-Link.

Κεφάλαιο 1. Διεργασία εξόρυξης πληροφορίας

Στο κεφάλαιο αυτό εισάγουμε τη διεργασία εξόρυξης πληροφορίας. Εξηγούμε την ιδέα ότι η εξόρυξη πληροφορίας δεν είναι οι λειτουργίες κάποιου μαγικού και μυστηριώδους μαύρου κουτιού το οποίο είναι κατανοητό μόνο από ένα μικρό αριθμό μάγων εξόρυξης, αλλά, στην πραγματικότητα, είναι μία διεργασία η οποία περιλαμβάνει όλα εκείνα τα τυπικά στοιχεία που ενυπάρχουν στα περισσότερα έργα τεχνολογίας της πληροφορικής. Υπάρχει μία τεράστια τάξη επιχειρηματικών προβλημάτων τα οποία μπορούν να επιλυθούν χρησιμοποιώντας μεθόδους εξόρυξης πληροφορίας ικανές να εκτελεστούν από απλούς θνητούς της Πληροφορικής. Θα εξετάσουμε μία διεργασία η οποία έχει εφαρμοστεί με επιτυχία σε πολλές περιπτώσεις επιχειρηματικών προβλημάτων. Η διεργασία αυτή παρουσιάζεται αρχικά σε πολύ υψηλό επίπεδο και στη συνέχεια συζητάμε κάθε κύριο συστατικό της.

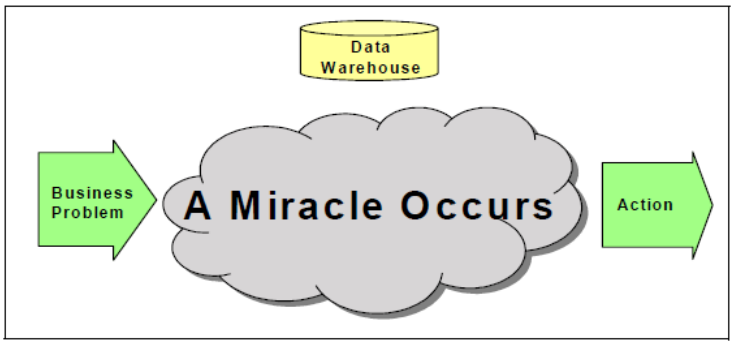
1. Απο-μυστικοποίηση της εξόρυξης πληροφορίας

Για πολλούς, ίσως για τους περισσότερους ανθρώπους, η εξόρυξη πληροφορίας είναι ένα παράξενο και σχεδόν μυστηριακό φαινόμενο. Είναι, ως εκ τούτου, ένα αντικείμενο όχι ξεκάθαρα κατανοητό. Αυτό οφείλεται στο γεγονός ότι μπορεί να περιλαμβάνει πολλούς αρκετά τεχνικούς μαθηματικούς αλγορίθμους και να απαιτεί λεπτομερή και πολύπλοκη ανάλυση των αποτελεσμάτων που παράγει. Κυρίως για αυτούς τους λόγους, η εξόρυξη πληροφορίας δεν έχει χρησιμοποιηθεί στο εύρος που θα ήταν δυνατό να χρησιμοποιηθεί. Κατά συνέπεια, η αξία που μπορεί να προσφέρει δεν έχει γίνει πλήρως αντιληπτή. Αυτό όμως σιγά σιγά αλλάζει. Με την ενσωμάτωση των διεργασιών εξόρυξης πληροφορίας εντός των συστημάτων εφαρμογών αλλά και την αποσαφήνιση του τρόπου λειτουργίας τους, η κατανόησή τους, συνεπώς και η χρήση τους, γίνονται ολοένα και πιο εύκολα.

2. Μία ιστορική ματιά στην εξόρυξη πληροφορίας

Αυτή η μυστηριακή αντίληψη μπορεί στην πραγματικότητα να έχει κάποια ιστορική βάση. Η εξόρυξη πληροφορίας γεννήθηκε στο πανεπιστημιακό περιβάλλον από θεωρητικούς με διδακτορία σε αντικείμενα όπως η στατιστική, που ανέπτυσσαν αλγόριθμους τους οποίους σήμερα κατατάσσουμε ως αλγόριθμους εξόρυξης πληροφορίας. Στις πρώιμες εφαρμογές εξόρυξης, οι αλγόριθμοι αυτοί κωδικοποιούνταν σε γλώσσες προγραμματισμού όπως η FORTRAN. Αυτές οι πρώιμες εφαρμογές ήταν κατά ανάγκη εστιασμένες στην υλοποίηση των αλγορίθμων και παρέμεναν ακαδημαϊκές στη φύση τους. Με το πέρασμα των χρόνων, διάφοροι κατασκευαστές ανέπτυξαν νέων επιπέδων εργαλεία ώστε να καταστήσουν την εφαρμογή των αλγορίθμων ευκολότερη. Αυτό είχε ως αποτέλεσμα τη δημιουργία ενός αριθμού αξιόλογων εργαλείων εξόρυξης. Εν τούτοις, η εξόρυξη πληροφορίας τείνει ακόμα να είναι μία ανεξάρτητη δραστηριότητα η οποία εκτελείται από εξειδικευμένο προσωπικό και η ορολογία αυτής παραμένει σε μεγάλο βαθμό ακαδημαϊκά προσανατολισμένη. Όλα αυτά καταλήγουν στο να διαφαίνεται αρκετά εσωτερική και κάτι τελείως ξεχωριστό από τα συνηθισμένα έργα πληροφορικής.

Μία κοινή αντίληψη της εξόρυξης πληροφορίας, από την χρήση της για την επίλυση ενός επιχειρηματικού προβλήματος αλλά και από την οπτική γωνία του προσωπικού πληροφορικής της επιχείρησης, απεικονίζεται στο σχήμα 1.1. Όταν προκύπτει κάποιο επιχειρηματικό πρόβλημα το οποίο πρέπει να επιλυθεί, το προσωπικό πληροφορικής ξεκινάει να εργάζεται πάνω σε αυτό. Οι μηχανικοί πληροφορικής κατανοούν πλήρως την αποθήκη δεδομένων που απεικονίζεται στην εικόνα, όταν όμως το επιχειρηματικό πρόβλημα δίνεται στην ομάδα εξόρυξης, φαίνεται ότι οι άνθρωποι της ομάδας αυτής φεύγουν μακρυά, επιτελούν κάποιο είδος θαύματος, και ξεπροβάλλουν από το σύννεφο με κάποιες αναλύσεις ή συστάσεις.

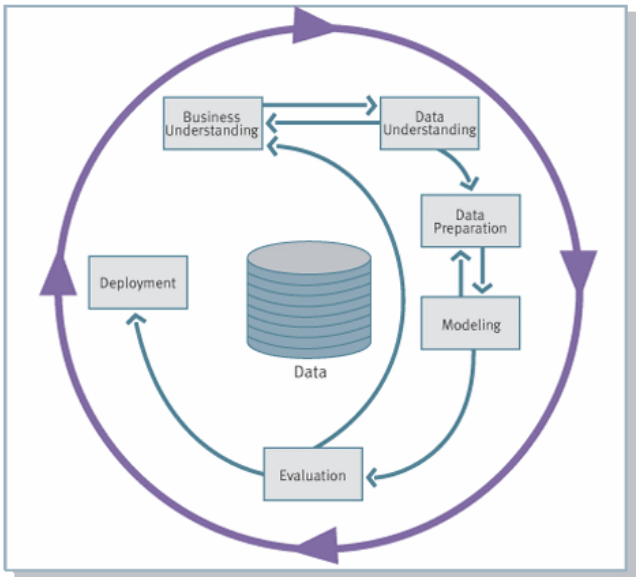


Σχήμα 1.1. Μία συνήθης λανθασμένη αντίληψη της εξόρυξης πληροφορίας.

Στην πορεία της εργασίας, ένα μεγάλο μέρος αυτού του σύννεφου θα εξαφανιστεί και θα μπορέσουμε έτσι να δούμε ξεκάθαρα τι πραγματικά συμβαίνει πίσω από αυτό. Πρέπει να σημειωθεί ότι η πλειοψηφία των εργασιών που εμπλέκονται στην ανάπτυξη μίας λύσης σε ένα επιχειρηματικό πρόβλημα χρησιμοποιώντας τεχνολογίες εξόρυξης πληροφορίας, είναι βασικά οι ίδιες με αυτές ενός οποιουδήποτε έργου πληροφορικής.

3. Το μοντέλο της διεργασίας εξόρυξης πληροφορίας

Το πρώτο πράγμα που πρέπει να αντιληφθούμε είναι ότι η εξόρυξη πληροφορίας δεν είναι μία αδιαπέραστη επικράτεια του σοφού και μυστηριώδους επιτελείου της, έτσι όπως διαφαίνεται. Στην πραγματικότητα, είναι πολύ όμοια με οποιοδήποτε έργο τεχνολογίας της πληροφορικής αφού είναι μία επαναληπτική διεργασία, το μεγαλύτερο μέρος της οποίας περιλαμβάνει τυπικές εργασίες πληροφορικής. Μόνο ένα μικρό κομμάτι της διεργασίας αυτής περιέχει την πραγματική τεχνολογία εξόρυξης πληροφορίας.



Σχήμα 1.2. Οι φάσεις της διεργασίας εξόρυξης πληροφορίας

Ο κύκλος ζωής ενός έργου εξόρυξης πληροφορίας αποτελείται από έξι φάσεις. Το παραπάνω σχήμα τις απεικονίζει στη διεργασία εξόρυξης. Η σειρά με την οποία απεικονίζονται δεν είναι αυστηρή. Η μετακίνηση προς τα πίσω ή εμπρός μεταξύ διαφορετικών φάσεων απαιτείται πάντα. Εξαρτάται από το αποτέλεσμα κάθε φάσης το ποια φάση ή ποια συγκεκριμένη εργασία μίας φάσης, θα πρέπει να εκτελεστεί στη συνέχεια. Τα βέλη υποδεικνύουν τις πιο σημαντικές και συχνές εξαρτήσεις μεταξύ των φάσεων.

Ο εξωτερικός κύκλος στο σχήμα 1.2 συμβολίζει την κυκλική φύση της εξόρυξης πληροφορίας. Η διεργασία δεν ολοκληρώνεται όταν η λύση την οποία επιδιώκουμε έχει αναπτυχθεί. Τα μαθήματα που αποκομίζουμε κατά τη διάρκεια εκτέλεσης της διεργασίας αλλά και αυτά από την ανάπτυξη της λύσης μπορούν να ενεργοποιήσουν

νέες, συχνά πιο ακριβής επιχειρηματικές ερωτήσεις. Μεταγενέστερες διεργασίες εξόρυξης επωφελούνται από τις εμπειρίες που αποκτώνται από τις προηγούμενες.

Παρακάτω, σκιαγραφούμε συνοπτικά την κάθε φάση.

1. **Κατανόηση της επιχείρησης**

   Η αρχική φάση εστιάζει στην κατανόηση των στόχων και απαιτήσεων του έργου από την προοπτική της επιχείρησης, μετατρέποντας τη γνώση αυτή σε ορισμό του προβλήματος εξόρυξης πληροφορίας και σε ένα προκαταρκτικό σχέδιο προορισμένο να επιτύχει τους στόχους του έργου.

2. **Κατανόηση των δεδομένων**

   Η φάση αυτή ξεκινάει με μία αρχική συλλογή δεδομένων και προβαίνει σε δραστηριότητες που αποσκοπούν στην εξοικείωση με τα δεδομένα, στην αναγνώριση προβλημάτων που αφορούν την ποιότητά τους, στην ανακάλυψη μιας πρώτης διορατικότητας εντός αυτών ή στον εντοπισμό ενδιαφερόντων υποσυνόλων τα οποία οδηγούν σε υποθέσεις περί κρυμμένης πληροφορίας.

3. **Προετοιμασία των δεδομένων**

   Η φάση προετοιμασίας των δεδομένων περιλαμβάνει δραστηριότητες για την κατασκευή του τελικού συνόλου δεδομένων (αυτό δηλαδή που θα τροφοδοτηθεί στο εργαλείο μοντελοποίησης) από τα αρχικά ακατέργαστα δεδομένα. Οι δραστηριότητες αυτές είναι πιθανό να εκτελεστούν πολλαπλές φορές χωρίς κάποια προκαθορισμένη σειρά. Μερικά παραδείγματα είναι η επιλογή πινάκων, εγγραφών και χαρακτηριστικών καθώς επίσης και η μετατροπή και ο καθαρισμός των δεδομένων.

4. **Μοντελοποίηση**

   Σε αυτή τη φάση, επιλέγονται και εφαρμόζονται διάφορες τεχνικές μοντελοποίησης με τις παραμέτρους τους να προσαρμόζονται στις ιδανικές τιμές. Τυπικά, υπάρχουν αρκετές τεχνικές για το ίδιο πρόβλημα εξόρυξης πληροφορίας. Μερικές τεχνικές έχουν συγκεκριμένες απαιτήσεις σε ότι αφορά τη μορφή των δεδομένων. Ως εκ τούτου, είναι συχνά αναγκαία η επιστροφή στη φάση προετοιμασίας των δεδομένων.

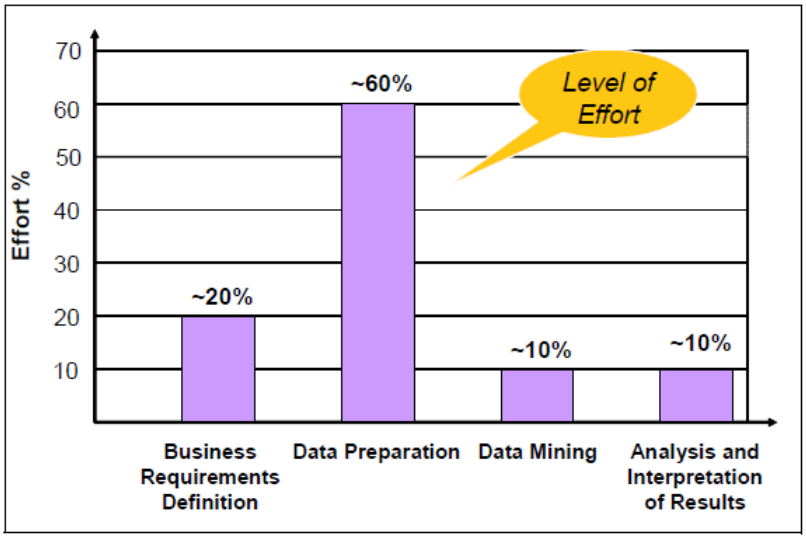
5. **Αξιολόγηση**

   Σε αυτό το στάδιο του έργου έχει ήδη χτιστεί ένα μοντέλο (ή πολλά μοντέλα) υψηλής ποιότητας από την οπτική γωνία της ανάλυσης δεδομένων. Προτού γίνει το επόμενο βήμα της τελικής ανάπτυξης του μοντέλου, είναι σημαντικό να αξιολογηθεί εξ ολοκλήρου και να επιθεωρηθούν τα βήματα που εκτελέστηκαν μέχρι την κατασκευή του, έτσι ώστε να υπάρχει η βεβαιότητα ότι επιτυγχάνει πλήρως τους επιχειρηματικούς στόχους. Βασική προτεραιότητα είναι να καθοριστεί αν υπάρχει κάποιο σημαντικό επιχειρηματικό πρόβλημα το οποίο δεν εξετάστηκε επαρκώς. Με την ολοκλήρωση αυτής της φάσης, θα πρέπει να έχει ληφθεί μία απόφαση σχετικά με τη χρήση των αποτελεσμάτων της εξόρυξης.

6. **Ανάπτυξη**

   Η δημιουργία του μοντέλου δε σηματοδοτεί και το τέλος του έργου. Ακόμη και αν ο σκοπός του μοντέλου είναι να αυξήσει τη γνώση των δεδομένων, θα πρέπει η γνώση αυτή να οργανωθεί και να παρουσιαστεί με τέτοιο τρόπο ώστε ο πελάτης να μπορεί να τη χρησιμοποιήσει. Η ανάπτυξη περιλαμβάνει συχνά την εφαρμογή "ζωντανών" μοντέλων στη διεργασία λήψης αποφάσεων ενός οργανισμού, για παράδειγμα την προσωποποίηση ιστοσελίδων σε πραγματικό χρόνο ή την επαναληπτική βαθμολόγηση (scoring) βάσεων δεδομένων μάρκετινγκ. Ωστόσο, ανάλογα με τις απαιτήσεις, η φάση ανάπτυξης μπορεί να είναι τόσο απλή όσο η δημιουργία μίας απλής αναφοράς ή τόσο πολύπλοκη όσο η υλοποίηση μίας επαναληπτικής διεργασίας εξόρυξης πληροφορίας εντός της επιχείρησης.

Ένα γράφημα το οποίο απεικονίζει τα συνήθη επίπεδα της προσπάθειας που πρέπει να καταβληθεί στις διάφορες φάσεις τις διεργασίας εξόρυξης, παρουσιάζεται στο ακόλουθο σχήμα.



Σχήμα 1.3. Τυπικά μεγέθη της συνολικής προσπάθειας που απαιτείται από τις φάσεις της διεργασίας εξόρυξης πληροφορίας

Οι τιμές που παρουσιάζονται είναι σχετικές και μπορεί να διαφέρουν από έργο σε έργο. Εκείνο το οποίο είναι αξιοσημείωτο είναι το μέγεθος της προσπάθειας που απαιτείται στη φάση προετοιμασίας των δεδομένων. Η προετοιμασία των δεδομένων είναι πολύ σημαντική καθώς οι ενέργειες που λαμβάνουν χώρα κατά τη διάρκειά της καθορίζουν σε μεγάλο βαθμό την ποιότητα των μοντέλων εξόρυξης. Όσο πιο καλά γίνει η προετοιμασία των δεδομένων, τόσο πιο σωστά μοντέλα θα παραχθούν στην αμέσως επόμενη φάση της μοντελοποίησης.

4. Η διεργασία που θα ακολουθήσουμε

Την ίδια ακριβώς διεργασία εξόρυξης θα ακολουθήσουμε για την αντιμετώπιση των ζητημάτων που τίθενται από τον σύνδεσμο HEAL-Link. Κάθε ένα κεφάλαιο που ακολουθεί αποτελεί μία ξεχωριστή φάση της διεργασίας αυτής. Σε κάθε φάση περιγράφουμε αναλυτικά τις ενέργειες που εκτελούμε και καταγράφουμε τα συμπεράσματα στα οποία καταλήγουμε.

Κεφάλαιο 2. Κατανόηση της επιχείρησης

Ο πρώτος στόχος του αναλυτή δεδομένων σε ένα έργο εξόρυξης πληροφορίας είναι να κατανοήσει, από επιχειρηματική σκοπιά, τι ακριβώς θέλει να επιτύχει η επιχείρηση. Συχνά, η επιχείρηση έχει πολλούς στόχους που ανταγωνίζονται ο ένας τον άλλον και θέτει πολλούς περιορισμούς που πρέπει να εξισορροπηθούν αποτελεσματικά. Στόχος του αναλυτή είναι να ανακαλύψει σημαντικούς παράγοντες οι οποίοι μπορούν να επηρεάσουν την έκβαση του έργου. Μία πιθανή συνέπεια παράλειψης αυτού του βήματος θα ήταν η διάθεση μεγάλης προσπάθειας για την παραγωγή σωστών απαντήσεων σε λάθος ερωτήματα.

Χρησιμοποιώντας τον όρο "επιχείρηση" από εδώ και στο εξής, θα εννοούμε τον σύνδεσμο HEAL-Link.

Σε αυτό το κεφάλαιο θα επιχειρήσουμε να καταγράψουμε τους επιχειρηματικούς στόχους, να προσδιορίσουμε τους στόχους εξόρυξης πληροφορίας, να αξιολογήσουμε την τρέχουσα κατάσταση και να προσδιορίσουμε τα εργαλεία και τις τεχνικές που θα χρησιμοποιήσουμε.

1. Ο σύνδεσμος HEAL-Link

1.1. Γενικές γνώσεις για τον σύνδεσμο HEAL-Link

Ο HEAL-Link (Hellenic Academic Libraries Link) είναι ο Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών που λειτουργεί υπό μορφή κοινοπραξίας και αποτελείται από τα ακόλουθα μέλη:

- 37 Ακαδημαϊκά Ιδρύματα (22 Πανεπιστήμια και 15 Τ.Ε.Ι.)
- 14 Ερευνητικά Ιδρύματα
- Ακαδημία Αθηνών
- Εθνική Βιβλιοθήκη Ελλάδος
- Βιβλιοθήκη της Βουλής
- Παιδαγωγικό Ινστιτούτο
- Εθνικό Ίδρυμα Αγροτικής Έρευνας
- Πανεπιστήμιο Κύπρου

Στόχοι του συνδέσμου είναι:

- Η συνεργασία, μέσω καθιέρωσης κοινής πολιτικής, στις συνδρομές των περιοδικών (έντυπων και ηλεκτρονικών) μεταξύ των μελών, με σκοπό την ορθολογική ανάπτυξη των συλλογών των περιοδικών μεταξύ των εταίρων, την εξοικονόμηση πόρων και την πρόσβαση σε μεγαλύτερο αριθμό πηγών για την κάλυψη των εκπαιδευτικών και ερευνητικών αναγκών των χρηστών των συμμετεχόντων Ιδρυμάτων.
- Η δημιουργία και λειτουργία Συλλογικού Καταλόγου βιβλιογραφικών εγγραφών (Union Catalogue) των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών και η χρήση των εγγραφών του Συλλογικού καταλόγου από κάθε μέλος της Κοινοπραξίας.
- Η από κοινού συνδρομή ηλεκτρονικών πηγών και υπηρεσιών πληροφόρησης, καθώς και δικαιωμάτων απομακρυσμένης πρόσβασης σε ηλεκτρονικές πηγές και υπηρεσίες πληροφόρησης, συμπεριλαμβανομένων των ηλεκτρονικών επιστημονικών περιοδικών.
- Η ανάπτυξη και καθιέρωση προτύπων για τις κάθε φύσης βιβλιοθηκονομικές εργασίες.

- Η μέριμνα για τη συνεχιζόμενη εκπαίδευση του προσωπικού των βιβλιοθηκών που είναι μέλη της.

- Η συνεργασία στη διάθεση του υλικού κάθε συμμετέχουσας βιβλιοθήκης μέσω διαδανεισμού και άλλων μεθόδων και πρακτικών που να εξασφαλίζουν και να διευκολύνουν τη διαθεσιμότητα του υλικού μεταξύ των εταίρων.

- Η συνεργασία με ανάλογους φορείς και οργανισμούς του εσωτερικού και του εξωτερικού για την εξασφάλιση της συμμετοχής της HEAL-Link στις διεθνείς εξελίξεις σε θέματα συνεργασίας βιβλιοθηκών και διαχείρισης πνευματικών δικαιωμάτων.

- Η ανάληψη κάθε άλλης πρωτοβουλίας που προάγει και αναπτύσσει τις Ακαδημαϊκές Βιβλιοθήκες της Ελλάδος μέσω κοινών δραστηριοτήτων και πρωτοβουλιών.

Ο HEAL-Link είναι μέλος στην International Coalition of Libraries Consortia, έναν διεθνή οργανισμό για την προώθηση των συνεργατικών προσπαθειών μεταξύ βιβλιοθηκών διεθνώς.

1.2. Σχετικά με τον διαδικτυακό τόπο HEAL-Link

Ο διαδικτυακός τόπος HEAL-Link (http://www.heal-link.gr/journals/) συνιστά κομβικό σημείο διαδικτυακής πρόσβασης της Ελληνικής και της Κυπριακής ακαδημαϊκής κοινότητας σε περισσότερους από 14.000 τίτλους επιστημονικών περιοδικών (2009), σε ηλεκτρονική/ψηφιακή μορφή. Λειτουργώντας επί μία δεκαετία περίπου, η βάση δεδομένων του HEAL-Link διαθέτει σήμερα εξαιρετικά πλούσια δεδομένα χρήσης (log files) του περιβάλλοντος. Στην πορεία, η διαδικτυακή εφαρμογή της βάσης δεδομένων του δικτυακού τόπου έχει εξελιχθεί/βελτιωθεί και ως προς τον κώδικά της και ως προς την έκδοση του συστήματος διαχείρισής της (IBM DB2). Σήμερα, τα δεδομένα του HEAL-Link υπόκεινται στο πλέον σύγχρονο είδος διαχείρισης και επεξεργασίας διαδικτυακού συστήματος το οποίο συμπεριλαμβάνει διακομιστή web (web server), διακομιστή εφαρμογής (application server) και διαχειριστή βάσεων δεδομένων (DBMS server).

Μέσω αυτής της διαδικτυακής πύλης τα μέλη του συνδέσμου έχουν πρόσβαση σε πλήρες κείμενο ηλεκτρονικών περιοδικών και βιβλίων και σε βιβλιογραφικές βάσεις δεδομένων.

Οι χρήστες του διαδικτυακού τόπου HEAL-Link χωρίζονται σε δύο κατηγορίες. Τους επισκέπτες και τα εγγεγραμμένα μέλη.

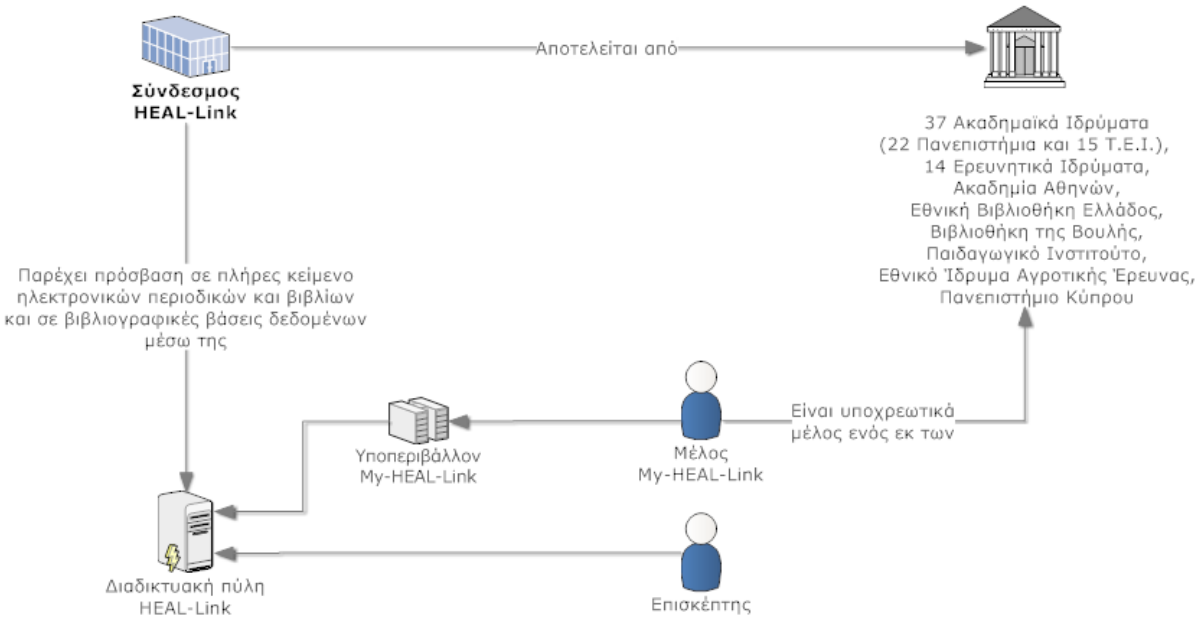
Επισκέπτης μπορεί να είναι οποιοσδήποτε εισάγει σε κάποιο πρόγραμμα φυλλομετρητή ιστού την ηλεκτρονική διεύθυνση http://www.heal-link.gr/journals/ και κατευθυνθεί στην διαδικτυακή πύλη του HEAL-Link.

Εγγεγραμμένα μέλη είναι χρήστες οι οποίοι χρησιμοποιούν την διαδικτυακή πύλη του HEAL-Link έχοντας προηγουμένως εγγραφεί (register) και εισαχθεί (login) στο υποπεριβάλλον My-HEAL-Link, που είναι μία υπηρεσία εξατομίκευσης.

Ο χρήστης έχει τη δυνατότητα να εγγραφεί και να αποκτήσει πρόσβαση στο υποπεριβάλλον My-HEAL-Link. Το συγκεκριμένο υποπεριβάλλον υποστηρίζει την εξατομικευμένη πρόσβαση στα ηλεκτρονικά περιοδικά της συλλογής. Μέσα από το My-HEAL-Link, ο χρήστης μπορεί να ενημερώσει τα προσωπικά του στοιχεία (Προσωπικά Στοιχεία) ή / και να διαμορφώσει το εξατομικευμένο περιβάλλον του (Προφίλ). Το δεύτερο μπορεί να γίνει επιλέγοντας περιοδικά που τον ενδιαφέρουν τα οποία θα εμφανίζονται αυτόματα στην εξατομικευμένη My-HEAL-Link σελίδα του. Επιπλέον, επιλέγοντας μέσα από το προφίλ “Αυτόματη ειδοποίηση”, ο χρήστης δηλώνει την επιθυμία να ειδοποιείται αυτόματα από το σύστημα κάθε φορά που η συλλογή των ηλεκτρονικών περιοδικών εμπλουτίζεται με ένα νέο περιοδικό που χαρακτηρίζεται από τους θεματικούς όρους που θα επιλέξει.

Για να εγγραφεί ένας χρήστης στο υποπεριβάλλον My-HEAL-Link θα πρέπει οπωσδήποτε να είναι μέλος ενός εκ των ιδρυμάτων μελών του συνδέσμου. Διαφορετικά, το σύστημα δε θα επιτρέψει την εγγραφή. Ο τρόπος με τον οποίο γίνεται ο έλεγχος των υποψήφιων προς εγγραφή χρηστών είναι με τη διεύθυνση e-mail που καταχωρούν. Εάν η κατάληξη της ηλεκτρονικής διεύθυνσης που εισάγει ο χρήστης κατά την εγγραφή του ανήκει σε κάποιο από τα ιδρύματα μέλη (π.χ. it.teithe.gr του ΑΤΕΙ Θεσσαλονίκης), τότε το σύστημα επιτρέπει την εγγραφή.

Στο παρακάτω σχήμα παρουσιάζεται ο σύνδεσμος HEAL-Link και οι συσχετίσεις του με τα ιδρύματα μέλη, την διαδικτυακή πύλη και τους χρήστες.



Σχήμα 2.1. Ο Σύνδεσμος HEAL-Link

1.2.1. Πού βρίσκονται αποθηκευμένα τα συγγράμματα

Είναι σημαντικό να διευκρινιστεί ότι τα ψηφιακά συγγράμματα (ηλεκτρονικά περιοδικά, βιβλία, κ.α.) τα οποία διατίθενται μέσω της διαδικτυακής πύλης HEAL-Link, δεν βρίσκονται αποθηκευμένα στην ίδια την πύλη, αλλά στους διαδικτυακούς τόπους των εκδοτών από τους οποίους εκδίδονται. Εκείνο το οποίο πραγματοποιεί ο διαδικτυακός τόπος HEAL-Link, είναι να συγκεντρώνει τους καταλόγους των εκδοτών με τους οποίους συνεργάζεται ο σύνδεσμος και να τους ενοποιεί σε έναν ενιαίο κατάλογο, δίνοντας έτσι τη δυνατότητα στα μέλη του συνδέσμου να περιηγούνται με μεγάλη ευκολία στα συγγράμματα πολλών διαφορετικών εκδοτών και να αποκτούν πρόσβαση σε αυτά.

Σχήμα 2.2. Ενιαίος κατάλογος ψηφιακών συγγραμμάτων

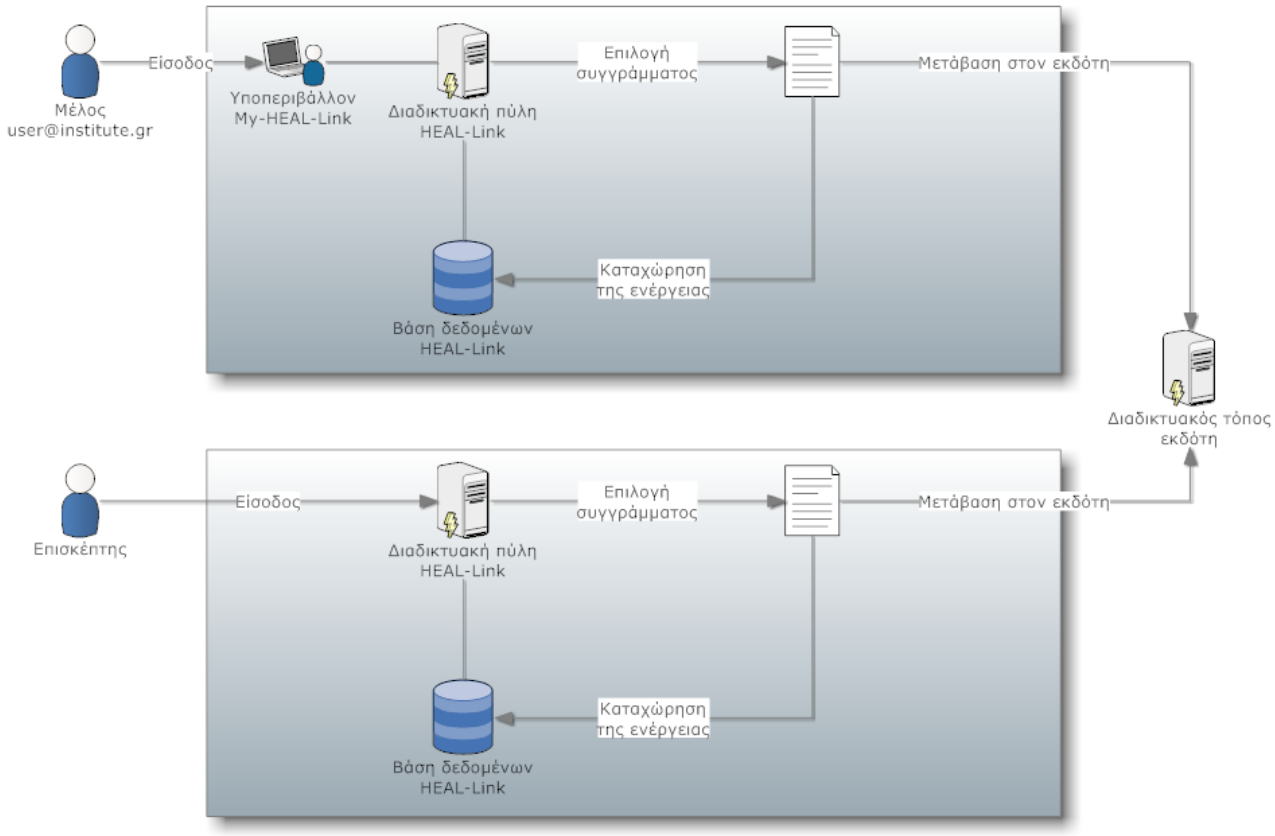
1.2.2. Στοχοποίηση και λήψη συγγραμμάτων

Όταν ο χρήστης εντοπίσει το σύγγραμμα που τον ενδιαφέρει και ζητήσει πρόσβαση στο περιεχόμενό του (κάνοντας κλικ στο όνομά του), τότε ο διαδικτυακός τόπος του HEAL-Link τον ανακατευθύνει στον διαδικτυακό τόπο του εκδότη από όπου θα μπορέσει να το μεταφορτώσει. Αυτό συμβαίνει ανεξάρτητα από το αν ο χρήστης είναι

επισκέπτης ή μέλος. Στο σημείο αυτό το έργο του HEAL-Link ολοκληρώνεται. Πλέον, αναλαμβάνει το αίτημα ο διαδικτυακός τόπος του εκδότη ο οποίος είναι υπεύθυνος να αποφασίσει εάν θα επιτρέψει την πρόσβαση του χρήστη στο σύγγραμμα ή όχι.

Κάθε δραστηριότητα επίσκεψης (στοχοποίησης) ψηφιακών συγγραμμάτων από τον κατάλογο του διαδικτυακού τόπου HEAL-Link καταγράφεται με λεπτομέρειες στην βάση δεδομένων του συστήματος. Οι πληροφορίες που καταγράφονται αφορούν τον χρήστη, την προέλευσή του, τη δραστηριότητα που εκτέλεσε, τη χρονική στιγμή της εκτέλεσης κ.α.

Στο ακόλουθο σχήμα αναπαρίσταται η λειτουργία του συστήματος όταν στοχοποιούνται συγγράμματα από εγγεγραμμένα μέλη και επισκέπτες.



Σχήμα 2.3. Στοχοποιήσεις συγγραμμάτων από εγγεγραμμένα μέλη και επισκέπτες

1.2.3. Πώς γίνεται ο έλεγχος πρόσβασης στα συγγράμματα

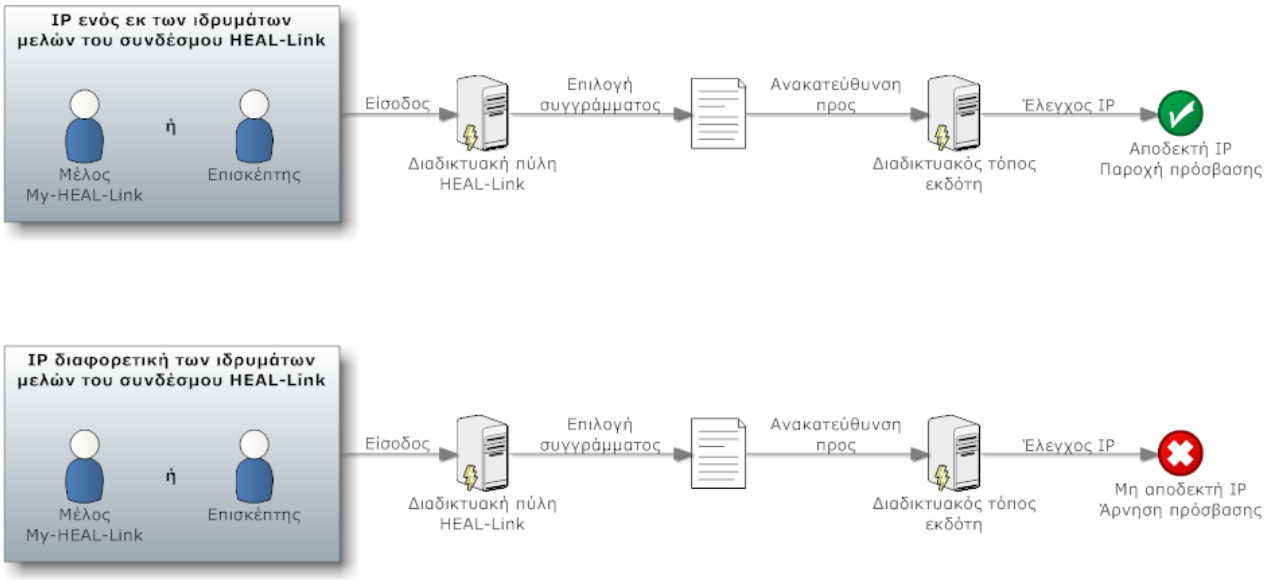
Ο έλεγχος πρόσβασης στα συγγράμματα γίνεται με αναγνώριση IP διευθύνσεων απευθείας από τους εκδότες χωρίς να εμπλέκεται σε αυτήν τη διαδικασία το σύστημα του HEAL-Link. Μόλις καταφθάσει ένα αίτημα λήψης ψηφιακού συγγράμματος στον διαδικτυακό τόπο ενός εκδότη, τότε ελέγχεται η IP προέλευσης του αιτήματος. Εάν η IP ανήκει σε κάποιο από τα ιδρύματα μέλη του συνδέσμου HEAL-Link, τότε ο εκδότης επιτρέπει την πρόσβαση στο σύγγραμμα. Διαφορετικά, το αίτημα απορρίπτεται.

Ο παραπάνω έλεγχος ισχύει και στις περιπτώσεις που ένας χρήστης θα ζητήσει ψηφιακό σύγγραμμα απευθείας από τον διαδικτυακό τόπο του εκδότη, χωρίς δηλαδή να έχει μεταβεί εκεί μέσω του HEAL-Link.

Ο διαχωρισμός των χρηστών σε επισκέπτες και εγγεγραμμένα μέλη που γίνεται από το HEAL-Link, είναι γνωστός μόνο στο ίδιο το σύστημα του HEAL-Link και όχι στους εκδότες. Αυτό σημαίνει ότι, ανεξάρτητα από το αν ο χρήστης που μετέβη στον διαδικτυακό τόπο ενός εκδότη μέσω του HEAL-Link είναι επισκέπτης ή εγγεγραμμένο μέλος (έχοντας προηγουμένως κάνει login), θα υποστεί την ίδια διαδικασία ελέγχου από τον εκδότη. Για παράδειγμα, αν ένας χρήστης εισέλθει στο υποπεριβάλλον My-HEAL-Link από μία IP διεύθυνση διαφορετική από αυτές των ιδρυμάτων μελών του συνδέσμου, και επιλέξει ένα σύγγραμμα, τότε ο εκδότης δεν θα του επιτρέψει την πρόσβαση στο σύγγραμμα γιατί δε θα αναγνωρίσει την IP διεύθυνση. Παρόλο δηλαδή που ο χρήστης είναι

μέλος του My-HEAL-Link, ο εκδότης δεν του επιτρέπει να λάβει το σύγγραμμα γιατί το αίτημά του προέρχεται από άγνωστη διεύθυνση IP.

Στο παρακάτω σχήμα συνοψίζεται η διαδικασία ελέγχου πρόσβασης στα συγγράμματα.



Σχήμα 2.4. Έλεγχος πρόσβασης στα συγγράμματα

1.2.4. Διαχείριση και λειτουργία της πύλης

Η διαδικτυακή πύλη σχεδιάστηκε και υλοποιήθηκε ως πτυχιακή εργασία από τους κ. Λεωνίδα Πισπιρίγγα, κ. Έλενα Σιδηροπούλου και κ. Υπάτιο Ασμανίδη, φοιτητές του τμήματος Πληροφορικής του Τεχνολογικού Εκπαιδευτικού Ιδρύματος (Τ.Ε.Ι.) Θεσσαλονίκης υπό την ακαδημαϊκή επίβλεψη του καθηγητή κ. Δ.Α. Δέρβου.

Η διαχείριση και λειτουργία της πύλης γίνεται στη Βιβλιοθήκη Φυσικής & Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, από τους κκ.:

- κα. Κλωντίνη Ξενίδου Δέρβου
- κα Πόπη Φλώρου
- κ. Βλάση Χατζησταύρου
- κ. Λεωνίδα Πισπιρίγγα

1.3. Κατανόηση των επιχειρηματικών στόχων

Βασικός στόχος των διαχειριστών είναι η επεξεργασία των δεδομένων χρήσης του διαδικτυακού τόπου HEAL-Link προς εξόρυξη πληροφορίας η σημασία/αξία της οποίας θα εξυπηρετεί δύο ανάγκες: α) τη διαδικασία λήψης αποφάσεων από το γραφείο διαχείρισης του διαδικτυακού τόπου HEAL-Link, και β) την εξυπηρέτηση των χρηστών του διαδικτυακού τόπου στην αναζήτηση επιστημονικής βιβλιογραφίας σχετικής προς το/α γνωστικό αντικείμενο/α της εξειδίκευσής τους. Ειδικά για την (β) ανάγκη, απαιτείται η ανάπτυξη πιλοτικού συστήματος λογισμικού αυτόματης παραγωγής συστάσεων (recommender system) όπου, π.χ., ανάλογα με τα περιοδικά ή/και θέματα που θα στοχοποιεί ένας χρήστης, το σύστημα θα του προτείνει και άλλα τα οποία άλλοι χρήστες έχουν στοχοποιήσει σε ανάλογου τύπου αναζητήσεις.

Αυτό θα έχει σαν αποτέλεσμα να δέχονται οι χρήστες απευθείας πληροφορία που τους ενδιαφέρει άμεσα, απαλλάσσοντάς τους από τον κόπο της αναζήτησης και καθιστώντας κατ'αυτόν τον τρόπο την υπηρεσία περισσότερο ελκυστική. Οι πληροφορίες αυτές ενδέχεται να μην έβγαιναν ποτέ στην επιφάνεια εάν έπρεπε από μόνοι τους, οι ίδιοι οι χρήστες, να τις αναζητήσουν.

Οι ερωτήσεις τις οποίες θέτουν οι διαχειριστές του διαδικτυακού τόπου είναι:

- Ποιους τίτλους συγγραμμάτων θεωρούν οι χρήστες ανάλογους;
- Ποιες κατηγορίες συγγραμμάτων τείνουν να επισκέπτονται μαζί;

1.4. Στόχοι εξόρυξης πληροφορίας

Σε κάθε επιχειρηματικό στόχο που εκφράζει μία απαίτηση της επιχείρησης, αντιστοιχεί ένας συγκεκριμένος στόχος εξόρυξης. Ο τελευταίος επιτυγχάνεται με μία συγκεκριμένη τεχνική εξόρυξης.

Αναλυτικά, για τους επιχειρηματικούς στόχους που διατυπώθηκαν παραπάνω, οι στόχοι εξόρυξης και οι αντίστοιχες τεχνικές έχουν ως εξής:

Πίνακας 2.1. Οι στόχοι της επιχείρησης μεταφρασμένοι σε στόχους εξόρυξης και οι αντίστοιχες τεχνικές εξόρυξης που θα εφαρμόσουμε

Στόχος επιχείρησης	Στόχος εξόρυξης	Τεχνική εξόρυξης
Ποιους τίτλους συγγραμμάτων θεωρούν οι χρήστες ανάλογους;	Ανακάλυψη κανόνων συσχετίσεων στις συναλλαγές των χρηστών του διαδικτυακού τόπου.	Κανόνες Συσχετίσεων
Ποιες κατηγορίες συγγραμμάτων τείνουν οι χρήστες να επισκέπτονται μαζί;	Ανακάλυψη κανόνων συσχετίσεων στις συναλλαγές των χρηστών του διαδικτυακού τόπου, συμπεριλαμβάνοντας πληροφορίες ταξινομίας.	Κανόνες Συσχετίσεων

2. Αξιολόγηση της τρέχουσας κατάστασης

2.1. Διαθέσιμο υλικό και λογισμικό

Οι πόροι υλικολογισμικού που θα χρησιμοποιήσουμε για την εκπόνηση της εργασίας καταγράφονται στον πίνακα 2.2.

Πίνακας 2.2. Πόροι υλικολογισμικού

	Hardware		Software	
	CPU	**RAM**	**Operating System**	**DBMS + Data Mining Tools**
Server	AMD Athlon 64 Dual Core 5600+	2GB	Windows 7 Professional 64-bit	IBM Infosphere Warehouse V9.7 (Data Server, Application Server, Client) 64-bit

2.2. Πηγές δεδομένων

Η βάση δεδομένων του διαδικτυακού τόπου HEAL-Link βρίσκεται εγκατεστημένη στη Βιβλιοθήκη Φυσικής & Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Η επεξεργασία των δεδομένων προς εξόρυξη πληροφορίας δε θα γίνει στην κεντρική βάση δεδομένων που τα φιλοξενεί. Τα δεδομένα θα εξαχθούν και στη συνέχεια θα μεταφερθούν και θα φορτωθούν σε κατάλληλο διακομιστή (στον διακομιστή που περιγράφεται παραπάνω στους πόρους υλικολογισμικού).

3. Αποτίμηση εργαλείων και τεχνικών

3.1. Εργαλεία

Το λογισμικό που θα χρησιμοποιήσουμε για την αποθήκευση των δεδομένων του διαδικτυακού τόπου HEAL-Link και την εφαρμογή τεχνικών εξόρυξης πληροφορίας είναι η πλατφόρμα της IBM, Infosphere Warehouse V9.7.

Η πλατφόρμα αποτελείται από τρεις διαφορετικές ομάδες λογισμικού (data server group, application server group, client group) η σύνθεση των οποίων δημιουργεί μία στιβαρή υποδομή για την αποθήκευση, επεξεργασία και ανάλυση μεγάλων όγκων δεδομένων καθώς επίσης και για την ανάπτυξη, εγκατάσταση και λειτουργία εφαρμογών που σχετίζονται με τα δεδομένα.

Τα συστατικά μέρη της πλατφόρμας IBM Infosphere Warehouse V9.7 φαίνονται στον παρακάτω πίνακα.

Πίνακας 2.3. Συστατικά μέρη της πλατφόρμας IBM Infosphere Warehouse V9.7

Ομάδα λογισμικού	Συστατικά μέρη

InfoSphere Warehouse Data Server Group	• DB2 Enterprise Server Edition
	  • Intelligent Miner®
	  • DB2 Query Patroller
	  • InfoSphere Federation Server Relational Wrappers

InfoSphere Warehouse Application Server Group	• Administration Console and Workload Manager
	  • SQL Warehousing (SQW) administration
	  • Cubing Services administration
	  • Intelligent Mining administration
	  • Workload Manager
	  • Unstructured Text Analysis
	• IBM Data Server Client
	• WebSphere® Application Server
	• Cubing Services cube server
	• Mining Blox®

InfoSphere Warehouse Client Group	• Design Studio
	  • SQL Warehousing (SQW) Tool
	  • Cubing Services modeling
	  • Intelligent Mining tools
	  • Unstructured Text Analysis tools
	  • Mining Blox tools
	• IBM Data Server client
	  • DB2 Query Patroller Center
	• Intelligent Miner Visualization
	• Cubing Services client
	• Administration Console Command Line Client

Τεχνικές πληροφορίες σχετικά με την πλατφόρμα Infosphere Warehouse V9.7 της IBM υπάρχουν στην ακόλουθη ηλεκτρονική διεύθυνση (δημόσια βιβλιοθήκη της IBM): http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/topic/com.ibm.dwe.navigate.doc/welcome_db2warehouse.html

Τα συστατικά μέρη της πλατφόρμας στα οποία θα επικεντρωθούμε και αυτά τα οποία θα μας απασχολήσουν περισσότερο, είναι:

DB2 Enterprise Server Edition: Το σύστημα διαχείρισης βάσεων δεδομένων που βρίσκεται στην καρδιά της πλατφόρμας Infosphere Warehouse. Το σύστημα αυτό θα αποτελέσει το χώρο αποθήκευσης και επεξεργασίας των δεδομένων του διαδικτυακού τόπου HEAL-Link.

Intelligent Miner: Πρόκειται για το σύστημα εξόρυξης πληροφορίας της πλατφόρμας. Ο Intelligent Miner συγκεντρώνει πλήθος αλγόριθμων εξόρυξης που καλύπτουν διάφορες τεχνικές εξόρυξης (clustering, associations, classification και prediction).

Intelligent Miner Visualizer: Το εργαλείο αυτό παρέχει δυνατότητα ανάλυσης των μοντέλων εξόρυξης που παράγει ο Intelligent Miner, οπτικοποιώντας τα αποτελέσματά τους σε ένα αλληλεπιδραστικό γραφικό περιβάλλον αναπτυγμένο σε Java. Παρέχει δυνατότητα προβολής και αξιολόγησης των αποτελεσμάτων.

Design Studio: Αποτελεί το βασικό περιβάλλον ανάπτυξης λύσεων επιχειρηματικής νοημοσύνης (BI) του Infosphere Warehouse. Το Design Studio ενσωματώνει σε ένα ενοποιημένο γραφικό περιβάλλον (βασισμένο στην πλατφόρμα Eclipse) τις ακόλουθες λειτουργίες:

1. Μοντελοποίηση φυσικών δεδομένων (physical data modeling)

2. Κατασκευή αποθηκών δεδομένων (DB2 SQL-based warehouse construction)

3. Δημιουργία μοντέλων κύβων OLAP (OLAP cube modeling)

4. Δημιουργία μοντέλων εξόρυξης πληροφορίας (data mining modeling)

3.2. Τεχνικές

Οι τεχνική εξόρυξης που θα χρησιμοποιήσουμε είναι η τεχνική των κανόνων συσχετίσεων.

Κανόνες συσχετίσεων

Οι κανόνες συσχετίσεων (association rules) αποτελούν μία σχετικά σύγχρονη μέθοδο για την εξαγωγή πληροφορίας από μεγάλες βάσεις δεδομένων. Εμφανίστηκαν για τις ανάγκες της ανάλυσης του "καλαθιού της αγοράς"[1]. Ο όρος αυτός προέρχεται από τις υπεραγορές (super-markets) στις οποίες ο καταναλωτής τοποθετεί σε ένα καλάθι το σύνολο των προϊόντων που επιθυμεί να αγοράσει. Οι υπεραγορές αυτές συγκεντρώνουν έναν τεράστιο όγκο πληροφοριών σχετικά με τις αγορές των πελατών τους, καθώς οι συναλλαγές κάθε πελάτη καταχωρούνται ηλεκτρονικά. Έτσι δημιουργήθηκε η ιδέα αξιοποίησης αυτής της πληροφορίας. Οι κανόνες συσχέτισης απλά εκφράζουν το αποτέλεσμα της ανάλυσης των χιλιάδων καλαθιών αγοράς των πελατών.

Ένας τέτοιος κανόνας είναι και ο εξής: "Οι πελάτες που αγοράζουν γάλα, αγοράζουν παράλληλα και ψωμί σε ποσοστό 60%". Ο παραπάνω κανόνας γράφεται σύντομα ως "γάλα → ψωμί (60%)". Η πρόταση αυτή παρουσιάζει ένα αίτιο, αγορά γάλακτος, και το συνδέει με ένα αποτέλεσμα, αγορά ψωμιού. Επίσης παρέχει μία ένδειξη για το πόσο πιθανό είναι να συμβαίνει μία τέτοια σχέση αιτίας-αιτιατού μέσω του ποσοστού που δίνεται. Οι κανόνες συσχέτισης επομένως, όπως υποδηλώνει το όνομά τους, είναι κανόνες "if-then" που συσχετίζουν αντικείμενα μεταξύ τους.

4. Επίλογος

Η κατανόηση της επιχείρησης και των στόχων της είναι ένα σημαντικό στάδιο στην εξέλιξη της διεργασίας εξόρυξης. Κατανοώντας τον τρόπο λειτουργίας της επιχείρησης και τα προβλήματα που θέτει, θα κατευθύνουμε την προσπάθειά μας στην εύρεση λύσεων για την καλύτερη αντιμετώπισή τους. Τα μοντέλα εξόρυξης που θα παράγουμε θα δίνουν λύσεις σε αυτά ακριβώς τα προβλήματα.

Έχοντας καταγράψει αναλυτικά τους στόχους του συνδέσμου HEAL-Link και τα μέσα που θα χρησιμοποιήσουμε για την ικανοποίησή τους, θα αναλύσουμε στο επόμενο κεφάλαιο με λεπτομέρειες την τεχνική των κανόνων

[1] Market Basket Analysis

συσχετίσεων και θα περιγράψουμε τους τρόπους με τους οποίους μπορούμε να παράγουμε κατανοητούς και πλούσιους σε γνώση κανόνες.

Κεφάλαιο 3. Παρουσίαση της τεχνικής των κανόνων συσχετίσεων

1. Η τεχνική

Η τεχνική των συσχετίσεων ανακαλύπτει συνδέσμους ή συσχετισμούς μεταξύ εγγραφών δεδομένων που αποτελούν τμήματα κάποιων μοναδικών γεγονότων που ονομάζονται συναλλαγές. Για παράδειγμα, εάν η συναλλαγή αναπαριστά το καλάθι μίας υπεραγοράς (super market), τότε οι εγγραφές δεδομένων αναπαριστούν τα αντικείμενα που περιέχονται στο καλάθι και αγοράζονται μαζί, π.χ. πατατάκια και μπύρα. Εάν η συναλλαγή αναπαριστά το επεισόδιο της αρρώστιας ενός ιατρικού ασθενή, τότε οι εγγραφές δεδομένων αναπαριστούν στοιχεία όπως τα συμπτώματα, τη φαρμακευτική αγωγή και τις αντιδράσεις του ασθενή στην περίθαλψη.

Η τεχνική των κανόνων συσχετίσεων απαντάει στο ερώτημα: Εάν ορισμένα αντικείμενα συμμετέχουν σε μία συναλλαγή, ποιο ή ποια άλλα αντικείμενα είναι πιθανό να συμμετέχουν στην ίδια συναλλαγή;

Ένα μοντέλο συσχετίσεων παράγει συνδυασμούς αντικειμένων που ονομάζονται στοιχειοσύνολα. Τα αντικείμενα αυτά (που αποτελούν τα περιεχόμενα των συναλλαγών) είναι οι μονάδες μεταξύ των οποίων αναγνωρίζονται οι συσχετίσεις. Κάθε στοιχειοσύνολο περιλαμβάνει ένα ή περισσότερα αντικείμενα και έχει μία σχετική συχνότητα εμφάνισης εντός του συνόλου των υπό ανάλυση συναλλαγών που ονομάζεται support. Οι τιμές του support χρησιμοποιούνται στη δημιουργία κανόνων οι οποίοι ποσοτικοποιούν τις συσχετίσεις μεταξύ αντικειμένων που συνυπάρχουν στις συναλλαγές. Οι κανόνες αυτοί εκφράζουν τις συσχετίσεις μεταξύ των αντικειμένων.

Οι κανόνες συσχετίσεων περιγράφονται με τους ακόλουθους όρους.

Για τον κανόνα [Γάλα] ==> [Δημητριακά]:

Σώμα του κανόνα:
: Ένα ή περισσότερα αντικείμενα σε μία συναλλαγή που συνεπάγονται την παρουσία ενός άλλου αντικειμένου. Στο παράδειγμα, το σώμα του κανόνα αποτελεί το πρώτο αντικείμενο (γάλα).

Κεφαλή του κανόνα:
: Ένα αντικείμενο του οποίου η παρουσία στη συναλλαγή συνεπάγεται από την παρουσία των αντικειμένων του σώματος του κανόνα. Στο παράδειγμα, την κεφαλή του κανόνα αποτελεί το δεύτερο αντικείμενο (δημητριακά).

Support (Υποστήριξη):
: Το ποσοστό όλων των συναλλαγών που περιλαμβάνουν μαζί τα αντικείμενα του σώματος και της κεφαλής.

Confidence (Εμπιστοσύνη):
: Η πιθανότητα να υπάρχει στη συναλλαγή το αντικείμενο της κεφαλής, εφόσον υπάρχουν σε αυτήν τα αντικείμενα (ή το αντικείμενο) του σώματος.

Lift:
: Βαθμός στον οποίο το confidence είναι μεγαλύτερο (ή μικρότερο) από το αναμενόμενο.

Έστω ότι σε μία υπεραγορά ισχύουν οι ακόλουθες συνθήκες:

- Τα δημητριακά συμμετέχουν στο 20% όλων των συναλλαγών.
- Τα δημητριακά συμμετέχουν στο 60% των συναλλαγών που περιέχουν γάλα.
- Το 3.7% όλων των συναλλαγών περιέχουν γάλα και δημητριακά μαζί.

Τότε, οι ιδιότητες του κανόνα [Γάλα] ==> [Δημητριακά] διαμορφώνονται ως εξής:

- Support = 3.7%
- Confidence = 60%

- Lift = 60% / 20% = 3

Η τιμή 3 του lift σημαίνει ότι τα δημητριακά είναι τρεις φορές πιο πιθανό να υπάρχουν σε συναλλαγές που περιέχουν γάλα παρά σε οποιαδήποτε άλλη συναλλαγή.

Ο παραπάνω κανόνας θα μπορούσε να διαβαστεί με τον ακόλουθο τρόπο:

"Πελάτες οι οποίοι αγοράζουν γάλα αγοράζουν και δημητριακά στο 60% των περιπτώσεων. Ο κανόνας αυτός επηρεάζει το 3.7% των συναλλαγών που εξετάστηκαν. Επιπλέον, η πιθανότητα να υπάρχουν δημητριακά στις συναλλαγές που περιέχουν γάλα, είναι τρεις φορές μεγαλύτερη από την πιθανότητα να υπάρχουν σε οποιαδήποτε άλλη συναλλαγή."

Με τις τιμές των Support, Confidence και Lift έχουμε τη δυνατότητα να αξιολογήσουμε τους κανόνες που παράγει ο αλγόριθμος εξόρυξης και να αποφασίσουμε ποιοι από αυτούς είναι σημαντικοί για το επιχειρηματικό πρόβλημα που αντιμετωπίζουμε.

1.1. Support

Το support ενός κανόνα συσχέτισης είναι το ποσοστό των συναλλαγών που περιλαμβάνουν όλα τα αντικείμενα του κανόνα (της κεφαλής και του σώματος). Το ποσοστό αυτό υπολογίζεται λαμβάνοντας υπόψιν όλες τις συναλλαγές και δείχνει πόσο συχνά εμφανίζονται μαζί τα αντικείμενα της κεφαλής και του σώματος του κανόνα στο σύνολο των συναλλαγών που εξετάζονται.

Υπολογίζεται πολύ απλά με τη διαίρεση a/b, όπου:

a
: Το πλήθος των συναλλαγών που περιλαμβάνουν όλα τα αντικείμενα του κανόνα (της κεφαλής και του σώματος)

b
: Το πλήθος όλων των συναλλαγών

1.2. Confidence

Το confidence ενός κανόνα συσχέτισης είναι ένα ποσοστό το οποίο δείχνει πόσο συχνά εμφανίζεται το αντικείμενο της κεφαλής του κανόνα σε όλες εκείνες τις συναλλαγές που περιλαμβάνουν τα αντικείμενα του σώματος.

Η τιμή του confidence υποδηλώνει την αξιοπιστία του κανόνα. Όσο πιο μεγάλο είναι, τόσο πιο συχνά βρίσκονται συσχετισμένα τα αντικείμενα της κεφαλής και του σώματος.

Το confidence υπολογίζεται με μία απλή διαίρεση m/n, όπου:

m
: Το πλήθος των συναλλαγών που περιλαμβάνουν τα αντικείμενα της κεφαλής και του σώματος

n
: Το πλήθος των συναλλαγών που περιλαμβάνουν τα αντικείμενα του σώματος

1.3. Lift

Με την τιμή του lift μπορεί να γίνει αξιολόγηση της σημαντικότητας του κανόνα. Το lift ενός κανόνα είναι ο λόγος του confidence προς το expected confidence του κανόνα.

Το expected confidence ενός κανόνα ορίζεται ως το γινόμενο του support της κεφαλής επί το support του σώματος διαιρούμενο με το support του σώματος του κανόνα.

Το confidence ορίζεται ως ο λόγος του support της κεφαλής και του σώματος προς το support του σώματος.

Το lift υπολογίζεται ως εξής:

lift = confidence / expected_confidence = confidence / (s(body) * s(head) / s(body)) = confidence / s(head)

όπου:

s(body)
: Το support του σώματος του κανόνα

s(head)
: Το support της κεφαλής του κανόνα

Το expected confidence είναι πανομοιότυπο με το support της κεφαλής του κανόνα.

Το lift μπορεί να πάρει τιμές μεταξύ του 0 και του άπειρο:

- Τιμή του lift μεγαλύτερη του 1 υποδηλώνει ότι το σώμα και η κεφαλή του κανόνα εμφανίζονται μαζί πιο συχνά από όσο αναμενόταν. Αυτό σημαίνει ότι η εμφάνιση του σώματος του κανόνα επηρεάζει θετικά την εμφάνιση της κεφαλής του κανόνα.
- Τιμή του lift μικρότερη του 1 υποδηλώνει ότι το σώμα και η κεφαλή του κανόνα εμφανίζονται μαζί λιγότερο συχνά από όσο αναμενόταν. Αυτό σημαίνει ότι η εμφάνιση του σώματος του κανόνα επηρεάζει αρνητικά την εμφάνιση της κεφαλής του κανόνα.
- Τιμή του lift κοντά στο 1 υποδηλώνει ότι το σώμα και η κεφαλή του κανόνα εμφανίζονται μαζί σχεδόν τόσο όσο αναμενόταν. Αυτό σημαίνει ότι η εμφάνιση του σώματος του κανόνα δεν έχει σχεδόν καμία επιρροή στην εμφάνιση της κεφαλής του κανόνα.

2. Απαιτήσεις από τα δεδομένα εξόρυξης

Η τεχνική των κανόνων συσχετίσεων απαιτεί από τα δεδομένα να βρίσκονται σε συγκεκριμένη διάταξη πριν ξεκινήσει η εξόρυξή τους. Η διάταξη αυτή ονομάζεται διάταξη δεδομένων συναλλαγών (transactional layout) και φαίνεται στον ακόλουθο πίνακα:

Πίνακας 3.1. Διάταξη δεδομένων συναλλαγών

Transaction ID	Item ID
101	45
101	57
101	29
102	5
102	33
103	89
103	94
103	18

Για την τεχνική των κανόνων συσχετίσεων, η διάταξη αυτή αποτελείται από έναν πίνακα με δύο στήλες, τον κωδικό συναλλαγής (transaction id) και τον κωδικό αντικειμένου (item id).

1. Οι συναλλαγές είναι γεγονότα ή καταστάσεις που λαμβάνουν χώρα σε συγκεκριμένες χρονικές στιγμές. Μία συναλλαγή μπορεί να περιέχει ένα μοναδικό αντικείμενο ή πολλά αντικείμενα (όπως προϊόντα, γεγονότα, καταστάσεις). Μία συναλλαγή μπορεί να προσδιοριστεί από ένα μοναδικό ακολουθιακό αναγνωριστικό, από μία ημερομηνία ή από μία χρονοσφραγίδα.

2. Τα αντικείμενα είναι τα στοιχεία των συναλλαγών. Μία συναλλαγή μπορεί να αποτελείται από ένα αντικείμενο ή από πολλά αντικείμενα. Για παράδειγμα, μία συναλλαγή αγοράς μπορεί να περιλαμβάνει ένα ή περισσότερα αντικείμενα (προϊόντα). Ένα περιστατικό βλάβης των τμημάτων μιας μηχανής (συναλλαγή) μπορεί να αποτελείται από τη βλάβη ενός τμήματος ή πολλών τμημάτων (αντικείμενα). Ένα επεισόδιο ασθενείας (συναλλαγή) μπορεί να αποτελείται από ένα σύμπτωμα ή από πολλά συμπτώματα (αντικείμενα).

Ο κωδικός που προσδιορίζει μοναδικά τις συναλλαγές μπορεί να είναι χαρακτήρας, ακέραιος αριθμός ή ημερομηνία και ώρα (χρονοσφραγίδα). Ο κωδικός που προσδιορίζει μοναδικά τα αντικείμενα μπορεί να είναι χαρακτήρας ή ακέραιος αριθμός.

Επιπλέον του πίνακα συναλλαγών, μπορούν να χρησιμοποιηθούν προαιρετικά και πίνακες που φέρουν πληροφορίες αντιστοίχισης ονομάτων σε κωδικούς αντικειμένων αλλά και πίνακες ταξινομίας.

3. Αντιστοίχηση ονομάτων (name mapping)

Τα δεδομένα του πίνακα συναλλαγών ενδέχεται να είναι κωδικοποιημένα με τέτοιο τρόπο ώστε να καθίσταται δύσκολη η κατανόησή τους. Για παράδειγμα, οι κωδικοί των αντικειμένων των συναλλαγών είναι πιθανό να αναπαρίστανται με αριθμούς. Αυτό έχει ως αποτέλεσμα οι κανόνες που θα παράγει ο αλγόριθμος εξόρυξης να είναι πολύ δύσκολο να διαβαστούν, όπως για παράδειγμα ο κανόνας [12] ==> [15]. Ο κωδικός 12 μπορεί να αντιπροσωπεύει το γάλα και ο κωδικός 15 τα δημητριακά.

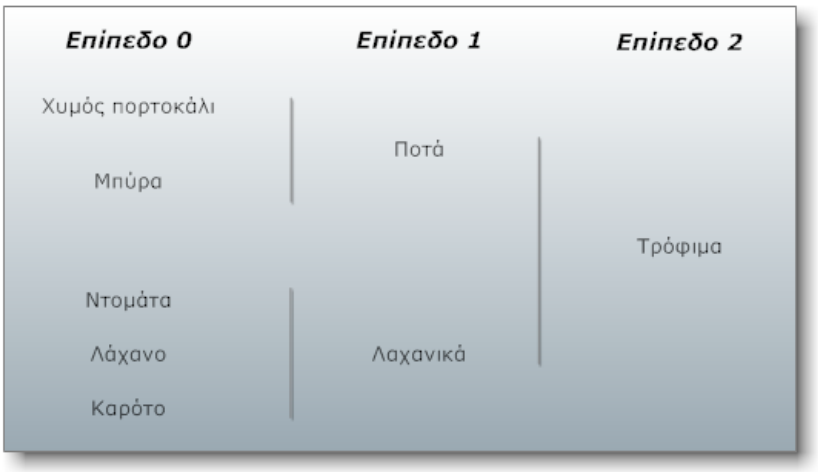
Ένας πίνακας αντιστοίχησης ονομάτων μπορεί να χρησιμοποιηθεί από τον Intelligent Miner για να αντικαταστήσει τους κωδικούς των αντικειμένων που συμμετέχουν στις συναλλαγές με τα ονόματά τους. Με αυτόν τον τρόπο, οι κανόνες που θα προκύψουν θα είναι πολύ πιο κατανοητοί. Αντί για κανόνες της μορφής [12] ==> [15], θα έχουμε κανόνες όπως [Γάλα] ==> [Δημητριακά].

Ο πίνακας αντιστοίχησης μπορεί να αποτελείται από δύο στήλες. Στη μία στήλη θα είναι αποθηκευμένοι οι κωδικοί και στην άλλη τα ονόματα με τα οποία θα αντικατασταθούν οι κωδικοί.

4. Ταξινομία (taxonomy)

Οι κανόνες συσχετίσεων που παράγει ο αλγόριθμος εξόρυξης μπορούν να γίνουν περισσότερο κατανοητοί εάν γίνει ομαδοποίηση των αντικειμένων των συναλλαγών σε κατηγορίες. Οι κατηγορίες αυτές μπορούν να ομαδοποιηθούν περαιτέρω σε υποκατηγορίες και ούτω καθεξής. Το αποτέλεσμα των ομαδοποιήσεων αυτών είναι μία ιεραρχία κατηγοριών με τα αντικείμενα των συναλλαγών στο χαμηλότερο επίπεδο. Η ιεραρχία αυτή ονομάζεται ταξινομία.

Το ακόλουθο σχήμα παρουσιάζει μία ταξινομία τριών επιπέδων:



Σχήμα 3.1. Παράδειγμα ταξινομίας τριών επιπέδων

Όταν εφαρμόζεται ταξινομία κατά την παραγωγή ενός μοντέλου συσχετίσεων, τότε οι συναλλαγές επεκτείνονται για τον υπολογισμό των κανόνων. Εάν το αντικείμενο μίας συναλλαγής ανήκει σε κάποια κατηγορία, τότε η κατηγορία αυτή προστίθεται στη συναλλαγή. Εάν η κατηγορία που προστέθηκε στη συναλλαγή ανήκει σε μία άλλη κατηγορία, τότε και η δεύτερη κατηγορία προστίθεται στη συναλλαγή και ούτω καθεξής. Κατά τη διάρκεια υπολογισμού των κανόνων, οι κατηγορίες που προστέθηκαν στις συναλλαγές αντιμετωπίζονται από τον αλγόριθμο με τον ίδιο ακριβώς τρόπο όπως και τα αντικείμενα. Κατά συνέπεια, οι κανόνες αναφέρονται όχι μόνο στα αντικείμενα αλλά και στις κατηγορίες. Για παράδειγμα, ο κανόνας:

[Χυμός πορτοκάλι] ==> [...]

μπορεί να συνοδεύεται από τους κανόνες:

[Ποτά] ==> [...]

[Τρόφιμα] ==> [...]

Ένα αντικείμενο ή μία κατηγορία μπορεί να είναι μέλος μίας ή περισσότερων κατηγοριών, ή καμίας κατηγορίας.

4.1. Χάρτες κατηγοριών (category maps)

Στον Intelligent Miner, ο ορισμός μίας ταξινομίας γίνεται χρησιμοποιώντας έναν ή περισσότερους χάρτες κατηγοριών. Χάρτης κατηγοριών είναι ένας πίνακας με τις ακόλουθες στήλες:

- Στήλη παιδιού
- Στήλη γονέα

Κάθε εγγραφή του πίνακα περιγράφει μία συσχέτιση μεταξύ ενός αντικειμένου και μίας κατηγορίας ή μεταξύ μίας κατηγορίας και μίας άλλης κατηγορίας. Το μέλος μίας κατηγορίας αποθηκεύεται στη στήλη παιδιού ενώ η ίδια η κατηγορία αποθηκεύεται στη στήλη γονέα.

Υπάρχουν δύο τύποι χαρτών κατηγοριών:

- Ένας *μη-αναδρομικός* χάρτης κατηγοριών μπορεί να διατηρεί συσχετίσεις μεταξύ δύο μόνο συνεχόμενων επιπέδων της ιεραρχίας κατηγοριών.
- Ένας *αναδρομικός* χάρτης κατηγοριών μπορεί να διατηρεί συσχετίσεις μεταξύ περισσότερων από δύο συνεχόμενων επιπέδων.

Για παράδειγμα, η ιεραρχία του σχήματος 3.1 μπορεί να αναπαρασταθεί με τους δύο μη-αναδρομικούς χάρτες κατηγοριών που φαίνονται παρακάτω:

Πίνακας 3.2. Χάρτης κατηγοριών 1: Μη-αναδρομικός

Παιδί	Γονέας
Χυμός πορτοκάλι	Ποτά
Μπύρα	Ποτά
Ντομάτα	Λαχανικά
Λάχανο	Λαχανικά
Καρότο	Λαχανικά

Πίνακας 3.3. Χάρτης κατηγοριών 2: Μη-αναδρομικός

Παιδί	Γονέας
Ποτά	Τρόφιμα
Λαχανικά	Τρόφιμα

Η ίδια ιεραρχία μπορεί να αναπαρασταθεί με τον ακόλουθο αναδρομικό χάρτη κατηγοριών:

Πίνακας 3.4. Χάρτης κατηγοριών: Αναδρομικός

Παιδί	Γονέας
Χυμός πορτοκάλι	Ποτά
Μπύρα	Ποτά
Ντομάτα	Λαχανικά

Παιδί	Γονέας
Λάχανο	Λαχανικά
Καρότο	Λαχανικά
Ποτά	Τρόφιμα
Λαχανικά	Τρόφιμα

5. Επίλογος

Χρησιμοποιώντας την τεχνική των κανόνων συσχετίσεων σε συνδυασμό με πληροφορίες αναζήτησης ονομάτων και ταξινομίας, θα μπορέσουμε να ανακαλύψουμε στα δεδομένα συναλλαγών του HEAL-Link συσχετίσεις μεταξύ συγγραμμάτων, θεματικών όρων, θεματικών υποκατηγοριών, θεματικών κατηγοριών και μεταξύ όλων των δυνατών συνδυασμών των παραπάνω. Οι κανόνες που θα προκύψουν θα συνοδεύονται από τους τρεις κυριότερους δείκτες ποιότητας, support, confidence και lift, τους οποίους θα μελετήσουμε και θα αξιολογήσουμε με σκοπό τον εντοπισμό των καλύτερων κανόνων.

Στο επόμενο κεφάλαιο εισάγουμε τα δεδομένα του διαδικτυακού τόπου HEAL-Link και ξεκινάμε την εξερεύνησή τους. Επιχειρούμε να κατανοήσουμε σε βάθος τον τρόπο λειτουργίας του διαδικτυακού συστήματος HEAL-Link και κάνουμε τις πρώτες επισημάνσεις των ζητημάτων που αφορούν την προετοιμασία των δεδομένων προς εξόρυξη.

Κεφάλαιο 4. Κατανόηση των δεδομένων

Στο κεφάλαιο αυτό εισάγουμε και εξερευνούμε τα δεδομένα της διαδικτυακής πύλης HEAL-Link. Περιγράφουμε πώς και από πού έγινε η εισαγωγή των δεδομένων στον Infosphere Warehouse Server, μελετάμε τα δεδομένα και επιλέγουμε εκείνα τα οποία θα χρησιμοποιήσουμε στη διεργασία εξόρυξης. Επίσης, επισημαίνουμε ορισμένα ζητήματα που αφορούν το μοντέλο δεδομένων και σχετίζονται με την προετοιμασία τους για εξόρυξη. Τέλος, αναλύουμε τα βασικότερα πεδία που θα αποτελέσουν ρόλους κλειδιά στη φάση εκτέλεσης του αλγόριθμου παραγωγής κανόνων συσχετίσεων (transaction id και item id).

1. Συγκέντρωση αρχικών δεδομένων

1.1. Εξαγωγή των δεδομένων από την πηγή τους

Αφού πρώτα εξάχθηκαν τα δεδομένα από την πηγή τους στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, στη συνέχεια μεταφορτώθηκαν στον ISW server όπου θα γίνει η επεξεργασία τους.

Η εξαγωγή των δεδομένων έγινε χρησιμοποιώντας την εντολή EXPORT[1] της DB2 η οποία τοποθετεί τα δεδομένα σε αρχεία της μορφής ixf[2]. Για τα δεδομένα κάθε πίνακα της βάσης δεδομένων δημιουργήθηκε ξεχωριστό αρχείο ixf. Επίσης, για κάθε πίνακα δημιουργήθηκε ξεχωριστό αρχείο κειμένου στο οποίο καταγράφηκαν μηνύματα τύπου log σχετικά με την πορεία της εξαγωγής. Για παράδειγμα, για τον πίνακα USER_PROFILE είχαμε το USER_PROFILE.ixf καθώς επίσης και το USER_PROFILE.txt.

Για τους 23 πίνακες που εξάχθηκαν από τη βάση δεδομένων, προέκυψαν από το πρόγραμμα εξαγωγής συνολικά 46 αρχεία (23 αρχεία τύπου ixf και 23 αρχεία καταγραφής). Στα παραπάνω προστέθηκε και ένα αρχείο sql με τον κώδικα DDL για τη δημιουργία της βάσης δεδομένων.

1.2. Εισαγωγή των δεδομένων στον ISW server

Όλα τα παραπάνω αρχεία συμπιέστηκαν και μεταφορτώθηκαν στον ISW server. Δημιουργήθηκε η βάση δεδομένων **HLDB** η οποία θα φιλοξενήσει τα δεδομένα του HEAL-Link και χρησιμοποιώντας κατάλληλα την εντολή LOAD[3] της DB2 έγινε η φόρτωσή τους στο σύστημα διαχείρισης βάσεων δεδομένων.

Παράδειγμα φόρτωσης των δεδομένων του πίνακα JOURNAL:

```
LOAD FROM "C:\HL_DATA\JOURNAL.ixf" OF IXF METHOD P (1, 2, 3, 4, 5, 6, 7)  MESSAGES ON
SERVER INSERT INTO DB2ADMIN.JOURNAL (J_ID, TITLE, LINK, CLASS, PUB_ID, DATE_ENTERED,
TIME_AVAILABLE) COPY NO INDEXING MODE AUTOSELECT ;
```

1.3. Δημιουργία έργου στο Design Studio για την επεξεργασία των δεδομένων και ενεργοποίηση της βάσης για εξόρυξη

Όλες οι εργασίες προετοιμασίας και εξόρυξης των δεδομένων θα εκτελεστούν στο περιβάλλον του Design Studio. Πριν από όλα θα πρέπει να προετοιμάσουμε κατάλληλα το περιβάλλον και να ενεργοποιήσουμε τη βάση δεδομένων HLDB για εξόρυξη.

1.3.1. Εκκίνηση του Design Studio και σύνδεση στη βάση δεδομένων HLDB

Ακολουθούμε τα παρακάτω βήματα.

[1]Παράδειγμα: db2 export to filename of ixf select * from table

[2](Από τη δημόσια βιβλιοθήκη της IBM) PC/IXF: PC version of the Integration Exchange Format (IXF), the preferred method for data exchange within the database manager. PC/IXF is a structured description of a database table that contains an external representation of the internal table.

[3]Παράδειγμα: db2 load from stafftab.ixf of ixf messages staff.msgs insert into userid.staff copy yes use tsm data buffer 4000

1. Εκκίνηση του Design Studio και ορισμός του χώρου εργασίας (workspace):

   - Επιλέγουμε κατά σειρά **Start > Programs > IBM InfoSphere Warehouse > ISWCOPY01 > Design Studio**.

   - Θα μας ζητηθεί να ορίσουμε τον χώρο εργασίας. Αφού εισάγουμε την επιθυμητή διαδρομή πατάμε **OK**. Για παράδειγμα:

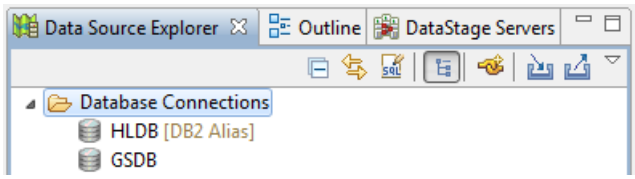
     ```
     C:\workspaces\miningtutorial
     ```

     Ο χώρος εργασίας αποτελεί μία κενή περιοχή στον δίσκο εντός της οποίας θα εργαστούμε. Εκεί θα αποθηκευτούν όλα τα αρχεία που θα παράγει το Design Studio.

2. Εμφάνιση του Data Source Explorer (εάν δεν είναι ήδη ορατός).

   - Επιλέγουμε **Window > Show View > Data Source Explorer**.

     Ο Data Source Explorer θα εμφανιστεί στην κάτω αριστερή γωνία του Design Studio.



Σχήμα 4.1. Data Source Explorer

Ο Data Source Explorer εμφανίζει τις συνδέσεις που έχουμε δημιουργήσει προς τις βάσεις δεδομένων. Θα χρειαστούμε μία σύνδεση προς τη βάση HLDB για τη δημιουργία του φυσικού μοντέλου δεδομένων.

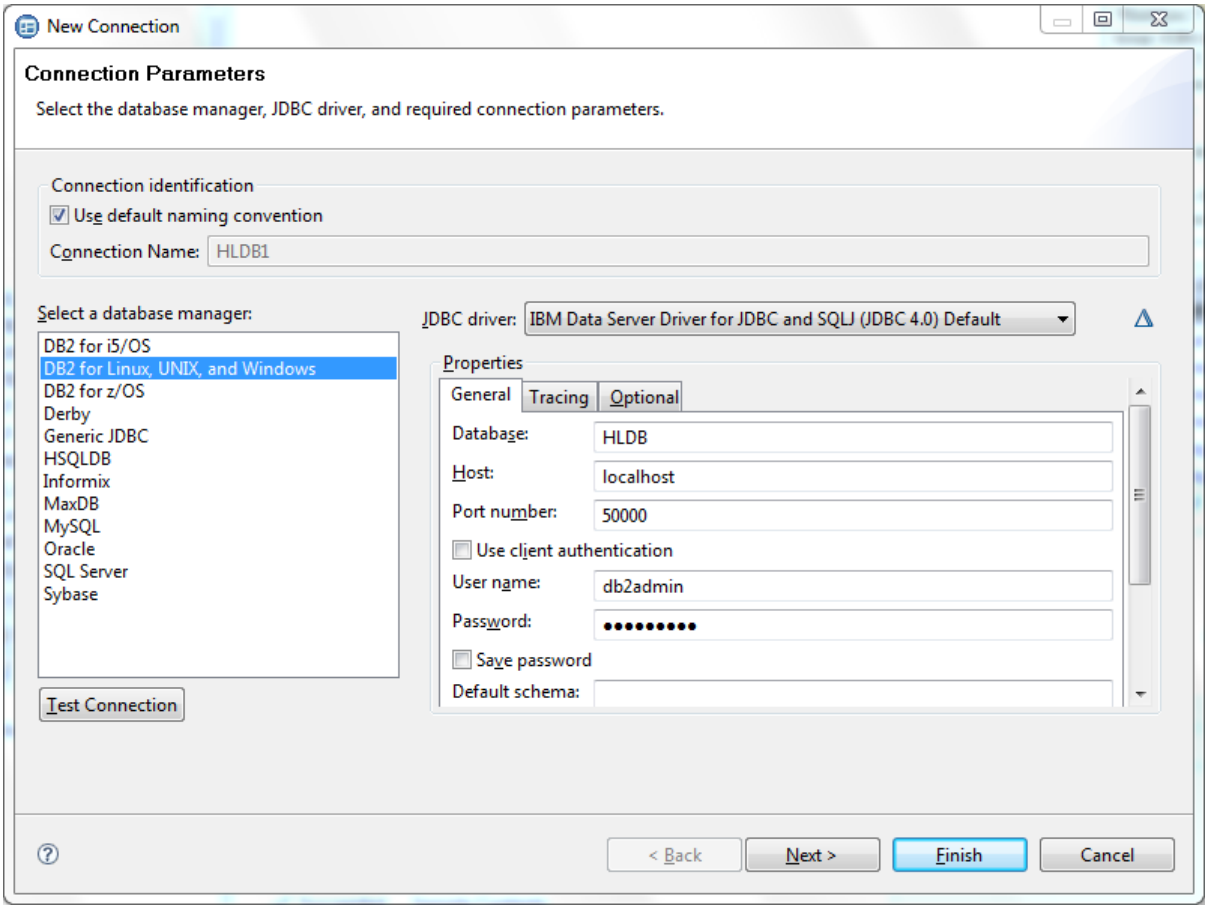
3. Δημιουργία σύνδεσης προς τη βάση δεδομένων HLDB.

   - Εκκίνηση του οδηγού δημιουργίας νέας σύνδεσης: Στον Data Source Explorer, δεξί κλικ στον φάκελο **Database Connections** και επιλογή **New**.

   - Συμπληρώνουμε τα πεδία του οδηγού δημιουργίας νέας σύνδεσης με βάση τις πληροφορίες του παρακάτω πίνακα:

Πίνακας 4.1. Πληροφορίες δημιουργίας νέας σύνδεσης προς τη βάση HLDB

Πεδίο	Τιμή
Select a database manager	DB2 for Linux, UNIX, and Windows
JDBC driver	Διατηρούμε την προεπιλογή.
Database	`HLDB`
Host	`localhost`
Port number	50000
Use client authentication	Μη επιλεγμένο
User ID	Εισάγουμε το όνομα χρήστη της βάσης δεδομένων HLDB. Για την εξόρυξη της βάσης ο χρήστης θα πρέπει να έχει δικαιώματα DB2 Administrator πάνω στη βάση.
Password	Εισάγουμε τον κωδικό του χρήστη.



Σχήμα 4.2. Στιγμιότυπο των ρυθμίσεων δημιουργίας σύνδεσης προς τη βάση δεδομένων HLDB

- Επιλέγουμε **Finish**. Η ρύθμιση της σύνδεσης ολοκληρώθηκε. Ταυτόχρονα συνδεόμαστε στη βάση δεδομένων.

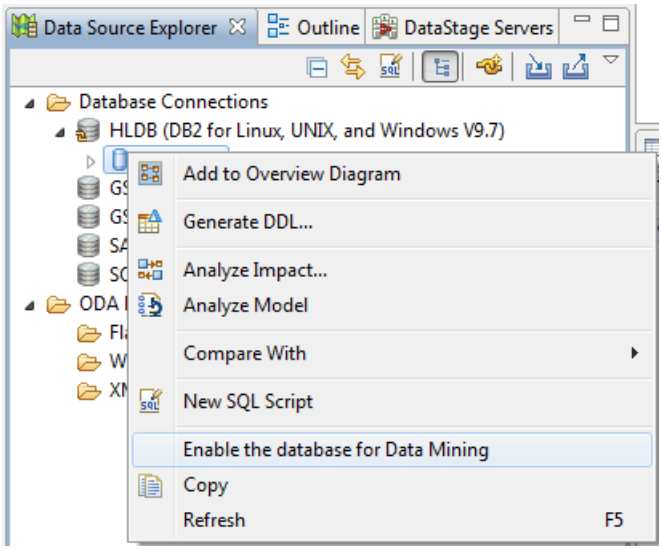
1.3.2. Ενεργοποίηση της βάσης δεδομένων HLDB για εξόρυξη

Για να εφαρμόσουμε τεχνικές εξόρυξης στη βάση δεδομένων θα πρέπει πρώτα να την ενεργοποιήσουμε για εξόρυξη. Ενεργοποιώντας μία βάση για εξόρυξη, προστίθενται σε αυτήν μέθοδοι εφαρμογής τεχνικών εξόρυξης (stored procedures). Η διαδικασία ενεργοποίησης έχει ως εξής:

1. Επεκτείνουμε το δέντρο **Database Connections** στον Data Source Explorer.

2. Επεκτείνουμε τη σύνδεση προς τη βάση δεδομένων **HLDB**.

3. Κάνουμε δεξί κλικ στη βάση δεδομένων **HLDB** και επιλέγουμε **Enable the database for Data Mining**.



Σχήμα 4.3. Ενεργοποίηση της βάσης δεδομένων προς εξόρυξη

Η βάση δεδομένων έχει πλέον ενεργοποιηθεί και είναι έτοιμη για εξόρυξη. Στη συνέχεια θα δημιουργήσουμε ένα νέο data warehousing project για την επεξεργασία των δεδομένων.

1.3.3. Δημιουργία νέου έργου Data Warehousing

Εκτελούμε τα ακόλουθα βήματα:

1. Επιλέγουμε **File > New > Data Warehousing Project**.

2. Εισάγουμε το όνομα του project:

```
HEAL-Link Data Warehousing Project
```

3. Επιλέγουμε **Finish**.

1.4. Τα αρχικά δεδομένα

Το συνολικό μέγεθος όλων των δεδομένων που φορτώθηκαν στον Infosphere Warehouse Server φτάνει περίπου τα 3.04GB.

Τα δεδομένα που λάβαμε από το HEAL-Link είναι μοιρασμένα σε 23 πίνακες. Οι πίνακες αυτοί μαζί με μία σύντομη περιγραφή παρουσιάζονται παρακάτω:

Πίνακας 4.2. Τα δεδομένα του διαδικτυακού τόπου HEAL-Link

Α/Α	Πίνακας	Πληροφορίες
1	AA_USER_PROFILE	Βοηθητικός πίνακας
2	CATEGORY	Οι θεματικές κατηγορίες στις οποίες είναι χωρισμένα τα συγγράμματα.
3	INSTITUTION	Τα ιδρύματα μέλη του συνδέσμου HEAL-Link.
4	J_SUBCAT	Συσχετισμοί των συγγραμμάτων με τις θεματικές υποκατηγορίες στις οποίες ανήκουν.
5	J_SUBJECT	Συσχετισμοί των συγγραμμάτων με τους θεματικούς όρους.
6	JOURNAL	Τα συγγράμματα.
7	JOURNAL_STATS	Στατιστικά που αφορούν τις στοχοποιήσεις των συγγραμμάτων.
8	NEWS	Νέα που αφορούν τον σύνδεσμο HEAL-Link.

Α/Α	Πίνακας	Πληροφορίες
9	NEWS_EN	Νέα που αφορούν τον σύνδεσμο HEAL-Link στα αγγλικά.
10	PROFILE_STATS	Στατιστικά που αφορούν το προφίλ των χρηστών.
11	PUBLISHER	Οι εκδότες με τους οποίους συνεργάζεται ο σύνδεσμος HEAL-Link.
12	REL_CAT	Βοηθητικός πίνακας
13	REL_CAT_MEMBER	Βοηθητικός πίνακας
14	REL_LINK	Βοηθητικός πίνακας
15	REL_TYPE	Βοηθητικός πίνακας
16	SESSIONS	Οι συνεδρίες.
17	SUBCATEGORY	Οι θεματικές υποκατηγορίες των συγγραμμάτων.
18	SUBJECT	Οι θεματικοί όροι των συγγραμμάτων.
19	SUBJECT_SUBCAT	Συσχετισμοί των θεματικών όρων με τις θεματικές υποκατηγορίες.
20	USER	Οι εγγεγραμμένοι χρήστες του υποπεριβάλλοντος My-HEAL-Link.
21	USER_PROFILE	Το προφίλ των εγγεγραμμένων χρηστών του υποπεριβάλλοντος My-HEAL-Link.
22	USER_STATS	Πληροφορίες που αφορούν τις ενέργειες των μελών του διαδικτυακού τόπου.
23	VISITS_STATS	Στατιστικά που αφορούν τις επισκέψεις χρηστών στον διαδικτυακό τόπο HEAL-Link.

2. Εξερεύνηση των δεδομένων

2.1. Απαιτήσεις από τα δεδομένα

Από το σύνολο των δεδομένων που φορτώθηκαν, θα πρέπει να γίνει κατάλληλη επιλογή εκείνων που είναι απαραίτητα για την υλοποίηση της τεχνικής εξόρυξης που θα εφαρμόσουμε. Δε θα γίνει χρήση όλων των δεδομένων παρά μόνο κάποιων υποσυνόλων που θα κριθούν αναγκαία.

Όπως αναφέρθηκε στο δεύτερο κεφάλαιο, στην αποτίμηση εργαλείων και τεχνικών, για την αντιμετώπιση των ζητημάτων του HEAL-Link θα χρησιμοποιήσουμε την τεχνική των κανόνων συσχετίσεων. Η τεχνική αυτή απαιτεί την ύπαρξη στα δεδομένα εκείνων των πληροφοριών οι οποίες συνθέτουν την έννοια της συναλλαγής (transaction).

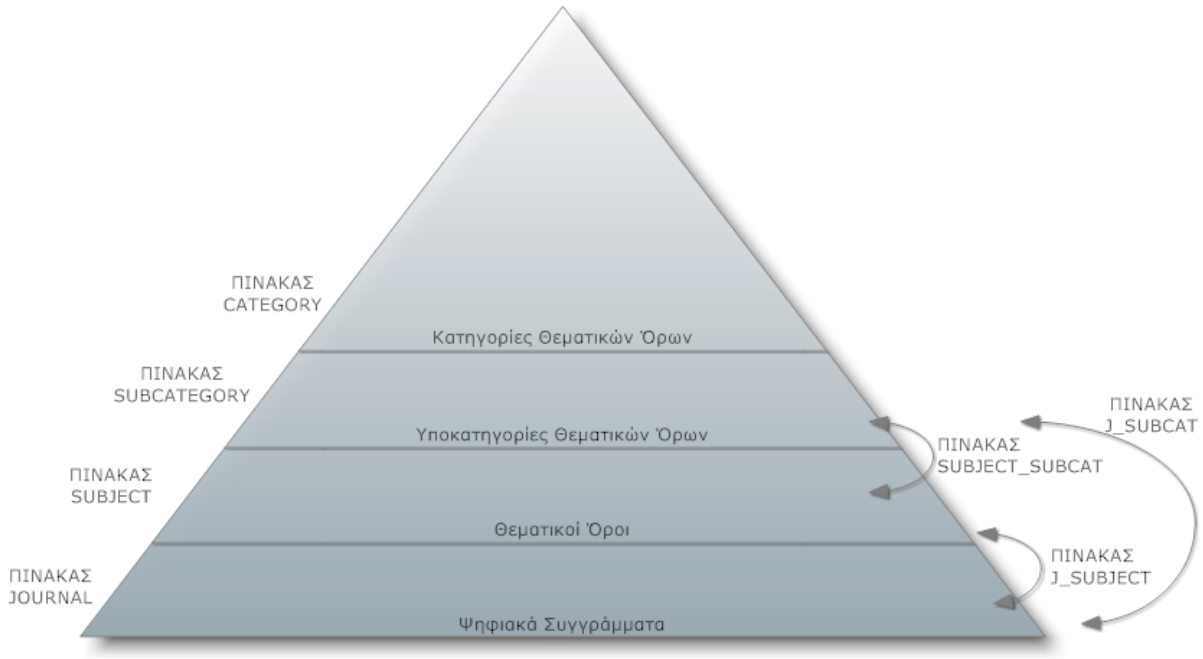
Τα δεδομένα που πρέπει πρωτίστως να αναζητήσουμε είναι αυτά που συνθέτουν τις συναλλαγές καθώς επίσης και τα αντικείμενα που αυτές περιλαμβάνουν. Επιπρόσθετα, πληροφορίες που αφορούν τα ψηφιακά συγγράμματα και την ιεραρχία κατηγοριών στην οποία κατατάσσονται κρίνονται αναγκαίες.

2.2. Είναι διαθέσιμες οι αναγκαίες πληροφορίες;

Το σύνολο των δεδομένων που συνθέτουν τις συναλλαγές του διαδικτυακού τόπου HEAL-Link βρίσκονται αποθηκευμένα στον πίνακα **JOURNAL_STATS**. Στον ίδιο πίνακα βρίσκονται και τα αντικείμενα των συναλλαγών (οι τίτλοι των στοχοποιημένων συγγραμμάτων).

Οι πληροφορίες που αφορούν και περιγράφουν τα συγγράμματα βρίσκονται στον πίνακα **JOURNAL**.

Ο τρόπος με τον οποίο αποθηκεύονται τα συγγράμματα στη βάση δεδομένων του HEAL-Link ακολουθεί μία συγκεκριμένη ιεραρχική δομή η οποία παρουσιάζεται στο ακόλουθο σχήμα:



Σχήμα 4.4. Η ιεραρχία των συγγραμμάτων

Στο κατώτερο επίπεδο βρίσκονται τα ψηφιακά συγγράμματα τα οποία περιγράφονται θεματικά από τα τρία επίπεδα που βρίσκονται παραπάνω.

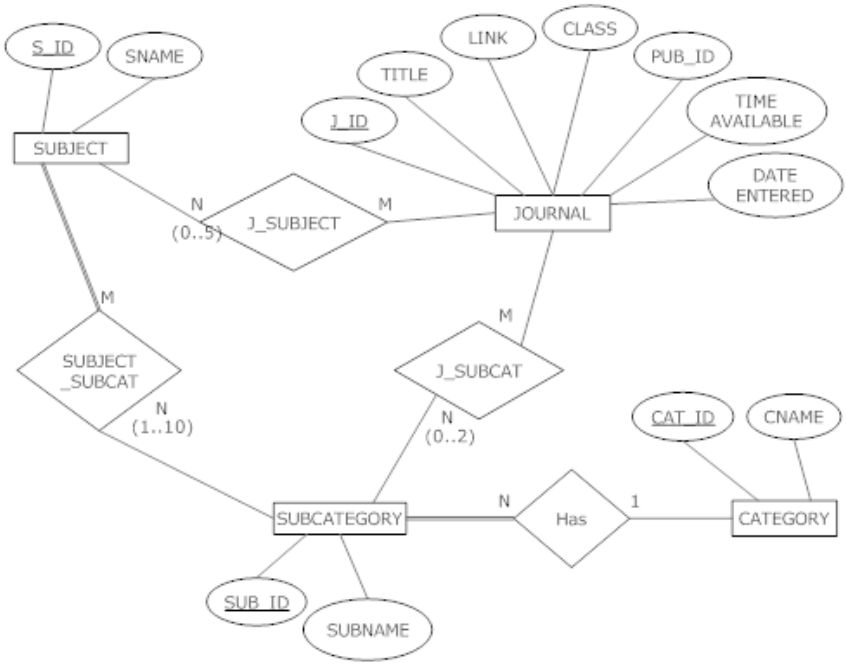
Κάθε σύγγραμμα μπορεί να έχει από έναν έως πέντε θεματικούς όρους που χαρακτηρίζουν το περιεχόμενό του. Μπορεί όμως να μην έχει και κανένα θεματικό όρο (ενδεχομένως γιατί εκκρεμεί η θεματική του ευρετηρίαση). Οι θεματικοί όροι στους οποίους κατατάσσονται τα συγγράμματα βρίσκονται στον πίνακα **SUBJECT**. Ο πίνακας που υλοποιεί την παραπάνω συσχέτιση είναι ο **J_SUBJECT**.

Κάθε σύγγραμμα μπορεί να είναι συσχετισμένο με μία ή δύο το πολύ θεματικές υποκατηγορίες. Μπορεί επίσης να μην είναι συσχετισμένο με καμία θεματική υποκατηγορία. Οι θεματικές υποκατηγορίες είναι αποθηκευμένες στον πίνακα **SUBCATEGORY**. Ο πίνακας που υλοποιεί τη συσχέτιση αυτή είναι ο **J_SUBCAT**.

Κάθε θεματικός όρος μπορεί να είναι συσχετισμένος με μία έως δέκα υποκατηγορίες θεματικών όρων. Ο πίνακας που υλοποιεί την τελευταία συσχέτιση είναι ο **SUBJECT_SUBCAT**.

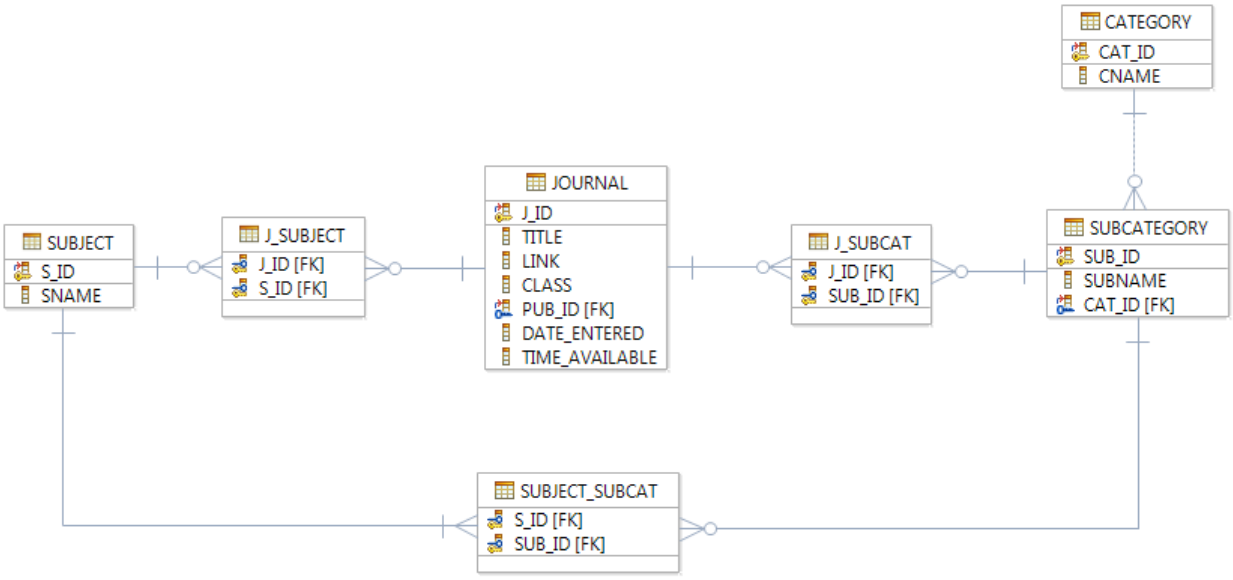
Τέλος, κάθε θεματική υποκατηγορία ανήκει υποχρεωτικά σε μία μόνο θεματική κατηγορία. Οι τελευταίες βρίσκονται στον πίνακα **CATEGORY**.

Στο ακόλουθο σχήμα απεικονίζεται το διάγραμμα οντοτήτων συσχετίσεων της ιεραρχίας των συγγραμμάτων και φαίνονται καθαρά οι συσχετίσεις μεταξύ των διαφόρων επιπέδων. Οι περιορισμοί πληθικότητας που βρίσκονται εντός παρενθέσεων ελέγχονται από το σύστημα διαχείρισης της βάσης.



Σχήμα 4.5. Διάγραμμα οντοτήτων συσχετίσεων της ιεραρχίας κατηγοριών των συγγραμμάτων

Η ίδια ιεραρχία στο φυσικό μοντέλο δεδομένων παρουσιάζεται στη συνέχεια.



Σχήμα 4.6. Φυσικό μοντέλο δεδομένων της ιεραρχίας κατηγοριών των συγγραμμάτων

Συνοψίζοντας:

- Κάθε σύγγραμμα μπορεί να συσχετίζεται με μηδέν, έναν ή περισσότερους (το πολύ πέντε) θεματικούς όρους (JOURNAL, J_SUBJECT, SUBJECT).
- Κάθε σύγγραμμα μπορεί να συσχετίζεται με μηδέν, μία ή περισσότερες (το πολύ δύο) θεματικές υποκατηγορίες (JOURNAL, J_SUBCAT, SUBCATEGORY).
- Κάθε θεματικός όρος συσχετίζεται υποχρεωτικά με μία ή περισσότερες (το πολύ δέκα) θεματικές υποκατηγορίες (SUBJECT, SUBJECT_SUBCAT, SUBCATEGORY).

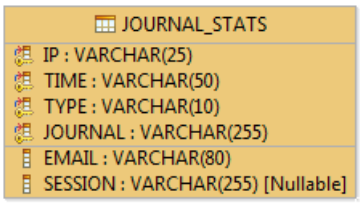
- Κάθε θεματική υποκατηγορία συσχετίζεται υποχρεωτικά με μία ακριβώς θεματική κατηγορία (SUBCATEGORY, CATEGORY).

2.3. Αναλυτική περιγραφή των επιλεγμένων δεδομένων

Παρακάτω περιγράφουμε αναλυτικά τη δομή των πινάκων που επιλέξαμε και τα δεδομένα τα οποία αποθηκεύουν.

2.3.1. JOURNAL_STATS

Ο πίνακας JOURNAL_STATS καταγράφει τις συναλλαγές που λαμβάνουν χώρα στον διαδικτυακό τόπο HEAL-Link και αφορούν την στοχοποίηση ψηφιακών συγγραμμάτων. Κάθε φορά που ένας χρήστης μεταβαίνει στον διαδικτυακό τόπο ενός συγγράμματος μέσω της ηλεκτρονικής διεύθυνσης του συγγράμματος (στήλη LINK του πίνακα JOURNAL), μία νέα γραμμή προστίθεται στον πίνακα JOURNAL_STATS η οποία περιλαμβάνει τις ακόλουθες πληροφορίες.



Σχήμα 4.7. Δομή του πίνακα JOURNAL_STATS

EMAIL — Το email του χρήστη. Πιθανές τιμές είναι το πραγματικό email του χρήστη εάν είναι μέλος του υποπεριβάλλοντος My-HEAL-Link και το αλφαριθμητικό "guest" εάν είναι επισκέπτης.

IP — Η διεύθυνση IP από την οποία ο χρήστης έφτασε στον διαδικτυακό τόπο.

TIME — Η χρονοσφραγίδα του συστήματος τη στιγμή καταχώρησης της νέας γραμμής στον πίνακα.

TYPE — Ο τρόπος με τον οποίο έφτασε ο χρήστης στο σύγγραμμα. Μερικά παραδείγματα περιλαμβάνουν: 'qsearch' αν χρησιμοποίησε τη λειτουργία γρήγορης αναζήτησης, 'alphabetic' αν χρησιμοποίησε την αλφαβητικά ταξινομημένη λίστα συγγραμμάτων, 'subject' αν χρησιμοποίησε την ιεραρχία "subject category" -> "subject sub-category" -> "subject" και 'selected' αν χρησιμοποίησε την προσωπική λίστα συγγραμμάτων που διατηρεί στο προφίλ του.

JOURNAL — Ο τίτλος του συγγράμματος.

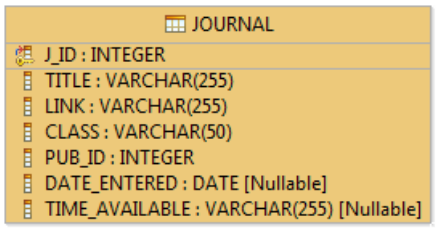
SESSION — Σε αυτό το πεδίο καταγράφεται η συνεδρία κατά τη διάρκεια της οποίας ο χρήστης εισήλθε στο σύστημα. Οι τιμές του πεδίου έχουν ως εξής: "ok" αν ο χρήστης είναι εγγεγραμμένο μέλος και κάνει login, "session_id" (μοναδικός αλφαριθμητικός κωδικός) αν ο χρήστης είναι επισκέπτης και NULL.

	EMAIL	IP	TIME	TYPE	JOURNAL	SESSION
2		66.249.66....	Jan 16, 2009 7:28:53 PM	alpha	Reactive and Functional Poly...	NULL
3	[redacted]@...	85.73.255....	Jan 16, 2009 7:28:54 PM	alpha	Sport in Society (formerly Cul...	ok
4	guest	147.52.89....	Jan 16, 2009 7:29:19 PM	alpha	Anesthesia	C9EEAB0DBB6...
5		66.249.66....	Jan 16, 2009 7:29:37 PM	alpha	Computer Animation and Vir...	NULL
6	[redacted]@...	85.73.255....	Jan 16, 2009 7:30:01 PM	alpha	Women___s Sports	ok
7	guest	195.251.1...	Jan 16, 2009 7:30:03 PM	alpha	Acta Psychiatrica Scandinavica	B4929A9705A8...
8		66.249.66....	Jan 16, 2009 7:30:20 PM	alpha	Functional	NULL
9	[redacted]@...	85.73.255....	Jan 16, 2009 7:30:57 PM	alpha	Scandinavian Journal of Medi...	ok
10		66.249.66....	Jan 16, 2009 7:31:06 PM	alpha	Algebras and Representation ...	NULL

Σχήμα 4.8. Δείγμα δεδομένων του πίνακα JOURNAL_STATS

2.3.2. JOURNAL

Ο πίνακας JOURNAL καταγράφει όλα τα ψηφιακά συγγράμματα που είναι διαθέσιμα μέσω του διαδικτυακού τόπου HEAL-Link. Για κάθε σύγγραμμα, φέρει τις ακόλουθες πληροφορίες:



Σχήμα 4.9. Δομή του πίνακα JOURNAL

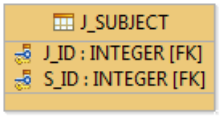
J_ID	Ένας μοναδικός κωδικός που ταυτοποιεί το σύγγραμμα.
TITLE	Ο τίτλος του συγγράμματος.
LINK	Ο σύνδεσμος URL που οδηγεί απευθείας στο σύγγραμμα στον διαδικτυακό τόπο του εκδότη.
CLASS	Κωδικός ταξινόμησης του συγγράμματος.
PUB_ID	Ο μοναδικός αριθμός ταυτοποίησης του εκδότη του συγγράμματος (ένα σύγγραμμα μπορεί να εκδίδεται από διαφορετικούς εκδότες σε διαφορετικές χρονικές περιόδους).
DATE_ENTERED	Η ημερομηνία κατά την οποία το σύγγραμμα έγινε διαθέσιμο μέσω της διαδικτυακής πύλης HEAL-Link.
TIME_AVAILABLE	Τα έτη κατά τη διάρκεια των οποίων έχουν πρόσβαση τα μέλη του συνδέσμου HEAL-Link στο κάθε σύγγραμμα βάση συμβολαίων.

	J_ID	TITLE	LINK	CLASS	PUB_ID	DATE_ENTERED	TIME_AVAILABLE
1	1	3C ON-LINE	http://portal.acm...	QA76.27	2	2003-01-01	(1994 - 1997)
2	2	AACN Clinical Issu...	http://gateway.o...	RT120 .I5A3	9	2003-01-01	(1993 - 2006)
3	3	Abacus	http://www.black...	HF5601 .A2	5	2003-01-01	(1997 -)
4	4	Abdominal Imaging	http://www.sprin...	QM543 .A2	12	2003-01-01	(1976 -)
5	5	About Campus	http://www3.inte...	LA226 .A2	14	2003-01-01	(1997 -)
6	6	ACC Current Jour...	http://www.scien...	RC666 .A25	6	2003-01-01	(1995 - 2005)
7	7	Access Control & ...	http://vnweb.hw...	TH9730	15	2003-01-01	(1999 - 2001)
8	8	Accident Analysis ...	http://www.scien...	HV675 .A...	6	2003-01-01	(1969 -)
9	9	Accident and Eme...	http://www.scien...	RT120 .E4...	6	2003-01-01	(1993 - 2007)
10	10	Accountability in ...	http://www.infor...	Q180.55	13	2003-01-01	(1989 -)

Σχήμα 4.10. Δείγμα δεδομένων του πίνακα JOURNAL

2.3.3. J_SUBJECT

Ο πίνακας J_SUBJECT υλοποιεί τη συσχέτιση πολλά προς πολλά μεταξύ των οντοτήτων JOURNAL και SUBJECT. Οι πληροφορίες που καταγράφει είναι ο κωδικός του συγγράμματος (J_ID) και ο κωδικός του θεματικού όρου (S_ID).



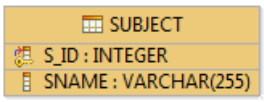
Σχήμα 4.11. Δομή του πίνακα J_SUBJECT

	J_ID	S_ID
1	1	579
2	1	1440
3	1	1441
4	2	2706
5	3	13
6	3	22
7	4	1
8	4	2
9	5	1911
10	6026	4045

Σχήμα 4.12. Δείγμα δεδομένων του πίνακα J_SUBJECT

2.3.4. SUBJECT

Ο πίνακας SUBJECT διατηρεί τους θεματικούς όρους στους οποίους κατατάσσονται τα συγγράμματα. Για κάθε θεματικό όρο καταχωρεί ένα μοναδικό κωδικό (S_ID) και το όνομα του όρου (SNAME).



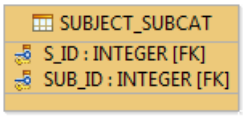
Σχήμα 4.13. Δομή του πίνακα SUBJECT

	S_ID	SNAME
1	1	Abdomen
2	2	Diagnostic imaging
3	3	Abnormalities, Human
4	4	Teratology
5	5	Academic libraries -- Australia
6	6	Research libraries -- Australia
7	7	Acarology
8	8	Ticks
9	9	Mites
10	10	Accidents

Σχήμα 4.14. Δείγμα δεδομένων του πίνακα SUBJECT

2.3.5. SUBJECT_SUBCAT

Ο πίνακας SUBJECT_SUBCAT υλοποιεί τη συσχέτιση πολλά προς πολλά μεταξύ των οντοτήτων SUBJECT και SUBCATEGORY. Οι πληροφορίες που καταγράφει είναι ο κωδικός του θεματικού όρου (S_ID) και ο κωδικός της θεματικής υποκατηγορίας (SUB_ID).



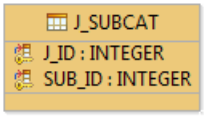
Σχήμα 4.15. Δομή του πίνακα SUBJECT_SUBCAT

	S_ID	SUB_ID
1	43	191
2	222	116
3	222	191
4	2706	119
5	13	26
6	22	26
7	1	79
8	2	79
9	1911	64
10	978	86

Σχήμα 4.16. Δείγμα δεδομένων του πίνακα SUBJECT_SUBCAT

2.3.6. J_SUBCAT

Ο πίνακας J_SUBCAT υλοποιεί τη συσχέτιση πολλά προς πολλά μεταξύ των οντοτήτων JOURNAL και SUBCATEGORY. Οι πληροφορίες που καταγράφει είναι ο κωδικός του συγγράμματος (J_ID) και ο κωδικός της θεματικής υποκατηγορίας (SUB_ID).



Σχήμα 4.17. Δομή του πίνακα J_SUBCAT

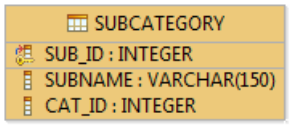
	J_ID	SUB_ID
1	2037	44
2	2	119
3	3	26
4	4	79
5	5	64
6	6026	52
7	7	18
8	3401	86
9	499	154
10	4683	130

Σχήμα 4.18. Δείγμα δεδομένων του πίνακα J_SUBCAT

2.3.7. SUBCATEGORY

Ο πίνακας SUBCATEGORY διατηρεί πληροφορίες για τις θεματικές υποκατηγορίες των συγγραμμάτων. Κάθε υποκατηγορία προσδιορίζεται μοναδικά από έναν κωδικό (SUB_ID), έχει ένα όνομα (SUBNAME) και αντιστοιχεί σε μία θεματική κατηγορία (CAT_ID). Το τελευταίο πεδίο (CAT_ID), είναι ο κωδικός που προσδιορίζει μοναδικά τη θεματική κατηγορία.

Στην περίπτωση αυτή έχουμε συσχέτιση ένα προς πολλά μεταξύ των οντοτήτων SUBCATEGORY και CATEGORY. Κάθε θεματική υποκατηγορία ανήκει σε μία μόνο θεματική κατηγορία ενώ σε κάθε κατηγορία μπορούν να αντιστοιχούν πολλές υποκατηγορίες.



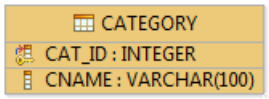
Σχήμα 4.19. Δομή του πίνακα SUBCATEGORY

	SUB_ID	SUBNAME	CAT_ID
1	1	Academies and learned societies	6
2	2	Aesthetics	20
3	3	Agriculture (General)	1
4	4	American literature	10
5	6	Anthropology	7
6	7	Aquaculture. Fisheries. Angling	1
7	8	Archaeology	2
8	9	Architecture	5
9	10	Armies: Organization, distribution, ...	18
10	11	Arts in general	5

Σχήμα 4.20. Δείγμα δεδομένων του πίνακα SUBCATEGORY

2.3.8. CATEGORY

Ο πίνακας CATEGORY καταγράφει πληροφορίες που αφορούν τις θεματικές κατηγορίες των συγγραμμάτων. Κάθε θεματική κατηγορία προσδιορίζεται μοναδικά από έναν κωδικό (CAT_ID) και έχει ένα όνομα (CNAME).



Σχήμα 4.21. Δομή του πίνακα CATEGORY

	CAT_ID	CNAME
1	1	Agriculture
2	2	Auxiliary sciences of history
3	3	Bibliography. Library science. Information resources
4	4	Education
5	5	Fine arts
6	6	General works
7	7	Geography. Anthropology. Recreation
8	8	History
9	9	History: America
10	10	Language and literature

Σχήμα 4.22. Δείγμα δεδομένων του πίνακα CATEGORY

2.3.9. Ποια χρονική περίοδο καλύπτουν τα δεδομένα;

Τα δεδομένα που θα εξετάσουμε καλύπτουν μία χρονική περίοδο περίπου τεσσάρων ετών. Η αρχαιότερη χρονοσφραγίδα εντός του πίνακα JOURNAL_STATS καταχωρήθηκε στις 7/6/2005 ενώ η νεότερη στις 9/3/2009. Συνολικά, στη συγκεκριμένη χρονική περίοδο καταχωρήθηκαν στον πίνακα JOURNAL_STATS **11.281.826** εγγραφές.

2.3.10. Ποσοτική ανάλυση

Στον ακόλουθο πίνακα γίνεται ποσοτική ανάλυση των επιλεγμένων δεδομένων. Οι πληροφορίες αυτές μας δίνουν μία γενική εικόνα του εύρους που καλύπτουν τα δεδομένα.

Πίνακας 4.3. Ποσοτική ανάλυση δεδομένων

Πίνακας	Πλήθος εγγραφών
JOURNAL	14.819
JOURNAL_STATS	11.281.826
J_SUBJECT	21.319
SUBJECT	7.531
SUBJECT_SUBCAT	10.023
J_SUBCAT	12.202
SUBCATEGORY	203
CATEGORY	20

2.4. Συγγράμματα που δεν ευρετηριάστηκαν θεματικά

Δεδομένου ότι η συμμετοχή των συγγραμμάτων στους πίνακες J_SUBJECT και J_SUBCAT δεν είναι υποχρεωτική, θα επιχειρήσουμε να εντοπίσουμε συγγράμματα τα οποία δε συμμετέχουν στις εν λόγω συσχετίσεις. Δηλαδή, συγγράμματα τα οποία είναι καταχωρημένα στη βάση δεδομένων αλλά δεν έχουν ευρετηριαστεί θεματικά.

Για τη συσχέτιση των συγγραμμάτων με τους θεματικούς όρους, κάθε σύγγραμμα μπορεί να συσχετίζεται με μηδέν, έναν ή περισσότερους (το πολύ πέντε) θεματικούς όρους.

- Υπάρχουν συγγράμματα τα οποία δε συσχετίζονται με κανένα θεματικό όρο;

```
select count(*)
from db2admin.JOURNAL
where j_id not in (select j_id from db2admin.J_SUBJECT);
```

  Ναι, 2900 συγγράμματα του πίνακα JOURNAL δε συσχετίζονται με κανένα θεματικό όρο.

Για τη συσχέτιση των συγγραμμάτων με τις υποκατηγορίες θεματικών όρων, κάθε σύγγραμμα μπορεί να συσχετίζεται με μηδέν, μία ή περισσότερες (το πολύ δύο) θεματικές υποκατηγορίες.

- Υπάρχουν συγγράμματα τα οποία δε συσχετίζονται με καμία θεματική υποκατηγορία;

```
select count(*)
from db2admin.JOURNAL
where j_id not in (select j_id from db2admin.J_SUBCAT);
```

  Ναι, 2900 συγγράμματα του πίνακα JOURNAL δε συσχετίζονται με καμία θεματική υποκατηγορία (είναι τα ίδια ακριβώς συγγράμματα που δε συσχετίζονται με κανένα θεματικό όρο).

Οι πληροφορίες αυτές μας βοηθούν να κατανοήσουμε τους κανόνες που θα παράγουν τα μοντέλα εξόρυξης που θα δημιουργήσουμε στη συνέχεια.

Συγκεκριμένα, όλα εκείνα τα συγγράμματα τα οποία δεν έχουν ευρετηριαστεί θεματικά δε θα είναι δυνατό να συμμετέχουν στην παραγωγή κανόνων που συσχετίζουν αντικείμενα διαφορετικών επιπέδων της ιεραρχίας συγγραμμάτων (π.χ. δε θα είναι δυνατή η συμμετοχή τους στην παραγωγή κανόνων όπως [Σύγγραμμα Α] ==> [Κατηγορία Β] ή [Θεματική Υποκατηγορία Α] ==> [Θεματική Υποκατηγορία Β]). Τα μη ευρετηριασμένα-θεματικά συγγράμματα θα χρησιμοποιηθούν από το μοντέλο εξόρυξης μόνο στην παραγωγή κανόνων συσχετίσεων που περιλαμβάνουν αντικείμενα του πρώτου επιπέδου (επίπεδο συγγραμμάτων). Δηλαδή, κανόνες της μορφής [Σύγγραμμα Α] ==> [Σύγγραμμα Β].

2.5. Επαναλαμβανόμενοι τίτλοι συγγραμμάτων

Επιχειρώντας να εξερευνήσουμε τα δεδομένα του πίνακα JOURNAL για να βεβαιωθούμε ότι κάθε σύγγραμμα είναι μοναδικό, εκτελούμε τον ακόλουθο κώδικα:

```
select title, count(*) as counts
from db2admin.JOURNAL
```

```
group by title
having count(*) > 1
order by counts desc;
```

Ομαδοποιούμε τις εγγραφές του πίνακα JOURNAL με βάση το πεδίο TITLE και ζητάμε να επιστραφούν εκείνα τα groups που έχουν πληθυσμό μεγαλύτερο από ένα. Δηλαδή, εκείνοι οι τίτλοι συγγραμμάτων που είναι καταχωρημένοι πάνω από μία φορά (αν υπάρχουν). Μερικά από τα αποτελέσματα που παίρνουμε είναι:

	TITLE	COUNTS
1	American Literary History	3
2	Compositio Mathematica	3
3	Early Music	3
4	Essays in Criticism	3
5	Hesperia	3
6	International Organization	3
7	Journal of the History of Medicine and Allied Scien...	3
8	Modern Judaism	3
9	Music and Letters	3
10	Opera Quarterly, The	3
11	Abstract and Applied Analysis	2
12	Adult Education Quarterly	2
13	Advances in Difference Equations	2
14	Advances in Physiology Education	2
15	Africa Today	2

Σχήμα 4.23. Τίτλοι συγγραμμάτων καταχωρημένοι στον πίνακα JOURNAL πάνω από μία φορά

Διαπιστώνουμε λοιπόν ότι υπάρχουν τίτλοι συγγραμμάτων που επαναλαμβάνονται έως και τρεις φορές. Κάθε φορά με διαφορετικό κωδικό J_ID. Συνολικά, το πλήθος των μοναδικών τίτλων συγγραμμάτων που επαναλαμβάνονται είναι 310. Οι επαναλήψεις αυτές μπορεί να οφείλονται σε δύο λόγους:

Πρώτον, στο γεγονός ότι ένα σύγγραμμα μπορεί να εκδίδεται από περισσότερους από έναν εκδότες. Κάθε επανάληψη αντιστοιχεί στο ίδιο σύγγραμμα εκδιδόμενο από διαφορετικό εκδότη. Για παράδειγμα, το περιοδικό "Hesperia" γίνεται διαθέσιμο μέσω του HEAL-Link από τρεις διαφορετικές πηγές: Project MUSE (29), ALPSP (28) και Wilson (15).

	J_ID	TITLE	LINK	CLASS	PUB_ID	DATE_ENTERED	TIME_AVAILABLE
1	13169	Hesperia	http://muse.jhu.edu/journal...	DS10 .H4	29	2007-02-13	(2005 only)
2	13168	Hesperia	http://www.swetswise.com/...	DS10 .H4	28	2007-02-13	(2002 -)
3	11037	Hesperia	http://vnweb.hwwilsonweb....	DS10 .H4	15	2006-06-29	(2004 -)

Σχήμα 4.24. Επαναλήψεις του περιοδικού Hesperia στον πίνακα JOURNAL

Ο δεύτερος λόγος είναι η περίπτωση όπου δύο διαφορετικά συγγράμματα έχουν τον ίδιο ακριβώς τίτλο.

Προκύπτει σε αυτό το σημείο το ερώτημα: Κάθε ξεχωριστό J_ID του ίδιου τίτλου στον πίνακα JOURNAL, ακολουθεί την ίδια διαδρομή στην ιεραρχία κατηγοριών των συγγραμμάτων (subject, subcategory, category); Με άλλα λόγια, όλα τα συγγράμματα του πίνακα JOURNAL που έχουν τον ίδιο τίτλο συσχετίζονται με τους ίδιους ακριβώς θεματικούς όρους, τις ίδιες θεματικές υποκατηγορίες και τις ίδιες θεματικές κατηγορίες;

Για να απαντήσουμε θα εκτελέσουμε τον ακόλουθο κώδικα (παράδειγμα περιοδικού "Hesperia"):

```
select j.j_id, j.title, p.pname, s.s_id, s.sname, sb.sub_id, sb.subname, c.cname
from db2admin.JOURNAL j,
  db2admin.SUBJECT s,
  db2admin.J_SUBJECT js,
  db2admin.SUBCATEGORY sb,
  db2admin.J_SUBCAT jsb,
  db2admin.CATEGORY c,
```

```
  db2admin.PUBLISHER p
where j.J_ID = js.J_ID and
   s.S_ID = js.S_ID and
   j.J_ID = jsb.J_ID and
   sb.SUB_ID = jsb.SUB_ID and
   sb.CAT_ID = c.CAT_ID and
   j.PUB_ID = p.PUB_ID and
   j.TITLE = 'Hesperia';
```

Κάνουμε JOIN τους πίνακες JOURNAL, SUBJECT, J_SUBJECT, SUBCATEGORY, J_SUBCAT, CATEGORY και PUBLISHER και προβάλουμε τα κατάλληλα πεδία για να ανακαλύψουμε τη διαδρομή που ακολουθούν στη θεματική ιεραρχία οι τρεις διαφορετικές επαναλήψεις του περιοδικού "Hesperia" (κωδικοί J_ID 13169, 13168 και 11037).

Τα επιστρεφόμενα αποτελέσματα είναι:

	J_ID	TITLE	PNAME	S_ID	SNAME	SUB_ID	SUBNAME	CNAME
1	13169	Hesperia	Project MUSE	6833	Greece -- Antiquities	60	History of Asia	History
2	13169	Hesperia	Project MUSE	6834	Greece -- Civilization -- To 146 B.C.	60	History of Asia	History
3	13169	Hesperia	Project MUSE	6835	Excavations (Archaeology) -- Greece	60	History of Asia	History
4	13168	Hesperia	ALPSP	6833	Greece -- Antiquities	60	History of Asia	History
5	13168	Hesperia	ALPSP	6834	Greece -- Civilization -- To 146 B.C.	60	History of Asia	History
6	13168	Hesperia	ALPSP	6835	Excavations (Archaeology) -- Greece	60	History of Asia	History
7	11037	Hesperia	Wilson	6833	Greece -- Antiquities	60	History of Asia	History
8	11037	Hesperia	Wilson	6834	Greece -- Civilization -- To 146 B.C.	60	History of Asia	History
9	11037	Hesperia	Wilson	6835	Excavations (Archaeology) -- Greece	60	History of Asia	History

Σχήμα 4.25. Η διαδρομή του περιοδικού "Hesperia" στη θεματική ιεραρχία

Σύμφωνα με το παραπάνω σχήμα, το περιοδικό "Hesperia" με κωδικό (J_ID) 13169 που διατίθεται από τον εκδότη "Project MUSE", συσχετίζεται με τρεις θεματικούς όρους:

1. Greece -- Antiquities

2. Greece -- Civilization -- To 146 B.C.

3. Excavations (Archaeology) -- Greece

Επιπλέον, ανήκει στην υποκατηγορία θεματικών όρων "History of Asia" η οποία ανήκει στην κατηγορία θεματικών όρων "History".

Για τις επαναλήψεις του περιοδικού Hesperia με κωδικούς 13168 και 11037 ισχύει η ίδια ακριβώς διαδρομή. Το μόνο που αλλάζει είναι ο εκδότης (ALPSP και Wilson).

Ακολουθεί ένα δεύτερο παράδειγμα με το περιοδικό "Music and Letters":

	J_ID	TITLE	PNAME	S_ID	SNAME	SUB_ID	SUBNAME	CNAME
1	5038	Music and Letters	Oxford University Press	3811	Music	99	Literature on music	Music and books on music
2	13497	Music and Letters	Project MUSE	3811	Music	99	Literature on music	Music and books on music
3	11526	Music and Letters	Wilson	3811	Music	99	Literature on music	Music and books on music

Σχήμα 4.26. Η διαδρομή του περιοδικού "Music and Letters" στη θεματική ιεραρχία

Στην περίπτωση αυτή οι τρεις επαναλήψεις του περιοδικού "Music and Letters" συσχετίζονται με ένα μόνο θεματικό όρο (τον θεματικό όρο "Music") και ανήκουν στην υποκατηγορία "Literature on Music" η οποία ανήκει στην κατηγορία "Music and books on music".

Μελετώντας τους 310 μοναδικούς τίτλους που επαναλαμβάνονται στον πίνακα JOURNAL, διαπιστώνουμε ότι οι επαναλήψεις των ίδιων τίτλων ακολουθούν πάντα την ίδια διαδρομή στη θεματική ιεραρχία. Η διαπίστωση αυτή θα μας φανεί πολύ χρήσιμη στη δημιουργία των μοντέλων εξόρυξης, κυρίως στα ζητήματα της αντιστοίχησης ονομάτων και της ταξινομίας.

Ο κώδικας που εκτελούμε είναι:

```
select j.j_id, j.title, p.pname, s.s_id, s.sname, sb.sub_id, sb.subname, c.cname
from db2admin.JOURNAL j,
  db2admin.SUBJECT s,
  db2admin.J_SUBJECT js,
  db2admin.SUBCATEGORY sb,
  db2admin.J_SUBCAT jsb,
  db2admin.CATEGORY c,
  db2admin.PUBLISHER p
where j.J_ID = js.J_ID and
   s.S_ID = js.S_ID and
   j.J_ID = jsb.J_ID and
   sb.SUB_ID = jsb.SUB_ID and
   sb.CAT_ID = c.CAT_ID and
   j.PUB_ID = p.PUB_ID and
   title in (select title
    from db2admin.JOURNAL
    group by title
    having count(*) > 1)
order by title;
```

2.6. Transaction ID και Item ID

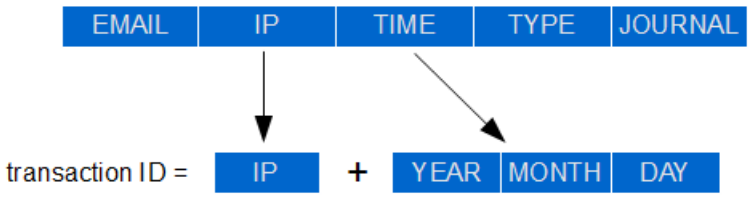
Όπως αναφέραμε ήδη στην παρουσίαση της τεχνικής των κανόνων συσχετίσεων, τα δεδομένα τα οποία θα τροφοδοτήσουμε στον αλγόριθμο εξόρυξης θα πρέπει να βρίσκονται σε διάταξη συναλλαγών (transaction id, item id). Πρέπει λοιπόν να εντοπίσουμε τα πεδία που θα αποτελέσουν το transaction id και το item id.

2.6.1. Τα πεδία SESSION και EMAIL ως Transaction ID

Σε προηγούμενη προσπάθεια που έγινε για την παραγωγή κανόνων συσχετίσεων από τα δεδομένα χρήσης του HEAL-Link, υπήρχε ένα βασικό πρόβλημα το οποίο κλήθηκε να αντιμετωπίσει η ομάδα εξόρυξης. Το πρόβλημα ήταν ποιο πεδίο από τον πίνακα JOURNAL_STATS θα χρησιμοποιηθεί ως transaction id. Δεδομένου ότι την περίοδο εκείνη ο διαδικτυακός τόπος του HEAL-Link δε χρησιμοποιούσε το σύστημα των συνεδριών (sessions) ώστε να προσδιορίζονται με ακρίβεια οι συναλλαγές, τα μέλη της ομάδας έπρεπε να εντοπίσουν και να χρησιμοποιήσουν ένα διαφορετικό transaction id.

Για την αντιμετώπιση του προβλήματος κατέληξαν σε δύο λύσεις οι οποίες περιγράφονται παρακάτω.

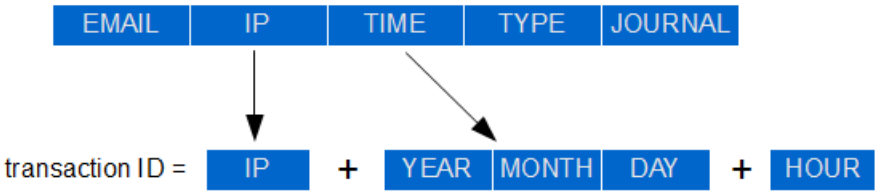
Ως πρώτη λύση, επέλεξαν να συνθέσουν το transaction id συνδυάζοντας το πεδίο IP με τα τμήματα YEAR, MONTH και DAY του πεδίου TIME, όπως φαίνεται στο παρακάτω σχήμα.



Σχήμα 4.27. Πρόταση Α για το transaction id

Αυτή η προσέγγιση έχει το εξής μειονέκτημα. Όλες οι επισκέψεις σε συγγράμματα που γίνονται από δύο ή περισσότερους χρήστες που μοιράζονται την ίδια διεύθυνση IP, την ίδια ημέρα, θεωρούνται ότι ανήκουν στην ίδια συναλλαγή. Η λύση αυτή προσεγγίζει αρκετά την πραγματικότητα όχι όμως τόσο όσο θα ήθελε η ομάδα. Γι'αυτό και προχώρησαν στη δεύτερη λύση.

Στη δεύτερη λύση, το transaction id θα αποτελεί ο συνδυασμός του πεδίου IP με τα τμήματα YEAR, MONTH, DAY και HOUR του πεδίου TIME. Με τον τρόπο αυτό περιορίζεται ακόμα περισσότερο η πιθανότητα να βρεθούν στην ίδια συναλλαγή δύο ή περισσότεροι χρήστες που μοιράζονται την ίδια διεύθυνση IP, καθώς για να συμβεί αυτό θα πρέπει να χρησιμοποιήσουν το σύστημα την ίδια ημέρα αλλά και την ίδια ώρα. Σίγουρα η δεύτερη λύση είναι πολύ πιο ασφαλής από την πρώτη.



Σχήμα 4.28. Πρόταση Β για το transaction id

Βεβαίως, θα μπορούσαμε να επεκταθούμε περαιτέρω και να προσθέσουμε στη σύνθεση του transaction id λεπτά ή και δευτερόλεπτα, ελαχιστοποιώντας κατά αυτόν τον τρόπο την πιθανότητα καταχώρησης μη ρεαλιστικών συναλλαγών.

Σε μεταγενέστερη έκδοση του συστήματος, το παραπάνω πρόβλημα λύθηκε ύστερα από την εισαγωγή μίας νέας πρακτικής, αυτής των συνεδριών (sessions). Ως συνεδρία θεωρείται μία ολοκληρωμένη περίοδος αλληλεπίδρασης ενός χρήστη με το σύστημα HEAL-Link. Ο τρόπος με τον οποίο λειτουργεί το σύστημα των συνεδριών όταν οι χρήστες είναι επισκέπτες διαφέρει από τον τρόπο λειτουργίας όταν οι χρήστες είναι εγγεγραμμένα μέλη.

Στη συνέχεια, παρουσιάζεται η συμπεριφορά του συστήματος ανάλογα με την κατηγορία των χρηστών.

2.6.1.1. Τιμές του πεδίου SESSION για τους επισκέπτες

Κάθε φορά που ένας επισκέπτης εισέρχεται στον διαδικτυακό τόπο του HEAL-Link, αυτομάτως το σύστημα του αντιστοιχεί ένα μοναδικό κωδικό που ονομάζεται **session id**. Ο κωδικός αυτός προσδιορίζει μοναδικά τη συνεδρία που μόλις ξεκίνησε και αντιστοιχήθηκε στον επισκέπτη. Η ζωή του session id διαρκεί όσο χρονικό διάστημα ο επισκέπτης βρίσκεται στην διαδικτυακή πύλη και αλληλεπιδρά με το σύστημα. Όταν αποσυνδεθεί, κλείνοντας για παράδειγμα τον web browser και περιμένοντας λίγο, η συνεδρία ολοκληρώνεται (το σύστημα αντιλαμβάνεται ότι ο χρήστης αποσυνδέθηκε και τερματίζει τη συνεδρία). Κατά τη διάρκεια της συνεδρίας, ο επισκέπτης μπορεί να ακολουθήσει τους συνδέσμους ενός ή περισσότερων συγγραμμάτων προς τους διαδικτυακούς τόπους των εκδοτών. Όλες αυτές οι επισκέψεις προς κάθε σύγγραμμα, θα καταγραφούν στον πίνακα JOURNAL_STATS κάτω από το ίδιο session id.

Είναι προφανές λοιπόν ότι κάθε συνεδρία αποτελεί και μία ξεχωριστή συναλλαγή. Συνεπώς, ο κωδικός που προσδιορίζει μοναδικά τη συνεδρία (session id) είναι και ο κωδικός της συναλλαγής (transaction id).

Οι διαχειριστές του συστήματος πρόσθεσαν στον πίνακα JOURNAL_STATS ένα νέο πεδίο με το όνομα SESSION, σκοπός του οποίου είναι να φιλοξενεί τους μοναδικούς κωδικούς των συνεδριών (session id). Το πεδίο SESSION είναι τύπου VARCHAR, δέχεται δηλαδή συμβολοσειρές. Οι κωδικοί των συνεδριών που παράγει το σύστημα του HEAL-Link έχουν μήκος 32 χαρακτήρων.

Κάθε φορά που ένας επισκέπτης ακολουθεί τον σύνδεσμο ενός συγγράμματος, μία νέα γραμμή προστίθεται στον πίνακα JOURNAL_STATS καταχωρώντας στο πεδίο SESSION το session id που προηγουμένως αντιστοίχησε το σύστημα στον επισκέπτη. Με αυτό τον τρόπο είναι φανερό ότι οι εγγραφές του πίνακα JOURNAL_STATS οι οποίες φέρουν την ίδια τιμή στο πεδίο SESSION, ανήκουν στην ίδια συναλλαγή.

2.6.1.2. Τιμές του πεδίου SESSION για τα εγγεγραμμένα μέλη

Εάν ο χρήστης είναι εγγεγραμμένο μέλος και εισέλθει στο σύστημα (κάνοντας login), τότε κάθε φορά που επισκέπτεται ένα σύγγραμμα μία νέα εγγραφή προστίθεται στον πίνακα JOURNAL_STATS καταχωρώντας στο πεδίο SESSION την τιμή '**ok**'. Στην περίπτωση αυτή δεν υπάρχει session id και κατ'επέκταση transaction id. Πρέπει λοιπόν να γίνει σύνθεση του transaction id συνδυάζοντας κάποια από τα πεδία του πίνακα JOURNAL_STATS.

Η προσέγγιση την οποία επιλέξαμε να ακολουθήσουμε είναι παρόμοια με τις δύο που περιγράψαμε παραπάνω. Επιλέξαμε να χρησιμοποιήσουμε ως transaction id το e-mail των μελών (πεδίο EMAIL) σε συνδυασμό με την ημέρα κατά τη διάρκεια της οποίας πραγματοποιήθηκαν οι στοχοποιήσεις (την ημέρα θα την εξάγουμε από το πεδίο TIMESTAMP). Αυτό σημαίνει ότι όλες οι εγγραφές του πίνακα JOURNAL_STATS που έχουν την ίδια τιμή στο πεδίο EMAIL και έχουν πραγματοποιηθεί την ίδια ημέρα (του ίδιου μήνα και του ίδιου έτους), θα ανήκουν στην ίδια συναλλαγή.

Είναι μία καλή προσέγγιση αν αναλογιστούμε ότι κατά τη διάρκεια μίας ημέρας, ένας συνηθισμένος χρήστης δε θα πραγματοποιήσει πολύ μεγάλο αριθμό στοχοποιήσεων αλλά ούτε και πολύ μικρό. Στις περισσότερες συναλλαγές το πλήθος των συμμετεχόντων αντικειμένων θα είναι αρκετά ικανοποιητικό ώστε να παρέχει γνώση σχετικά με το ποια συγγράμματα τείνουν τα μέλη να επισκέπτονται μαζί. Συνεπώς, δε χρειάζεται να περιορίσουμε περισσότερο το πλήθος των αντικειμένων κάθε συναλλαγής προσθέτοντας στο transaction id πεδία όπως η ώρα και τα λεπτά.

Επίσης, η συμμετοχή του e-mail στο transaction id διασφαλίζει ότι δεν υπάρχει κίνδυνος να προκύψουν συναλλαγές που να έχουν εκτελεστεί από δύο ή περισσότερους διαφορετικούς χρήστες.

2.6.1.3. Τιμές του πεδίου EMAIL για τους επισκέπτες

Εκτός από το πεδίο SESSION, το σύστημα του HEAL-Link καταχωρεί διαφορετικές τιμές και στο πεδίο EMAIL ανάλογα με την κατηγορία στην οποία ανήκει ο χρήστης (επισκέπτης, μέλος).

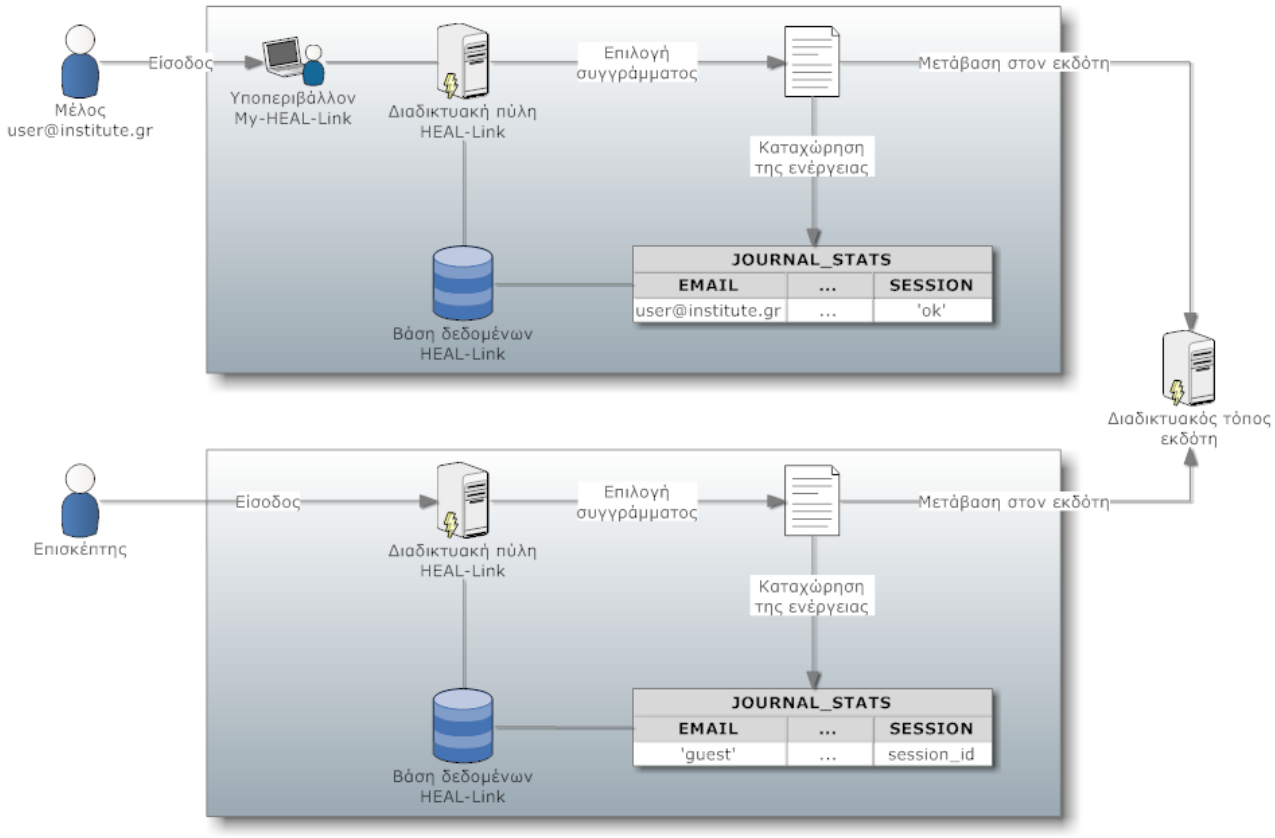
Κάθε φορά που ένας επισκέπτης ακολουθεί τον σύνδεσμο ενός συγγράμματος προς τον διαδικτυακό τοπο του εκδότη, τότε στη νέα γραμμή που προστίθεται στον πίνακα JOURNAL_STATS, στο πεδίο EMAIL, καταχωρείται το αλφαριθμητικό 'guest'.

2.6.1.4. Τιμές του πεδίου EMAIL για τα εγγεγραμμένα μέλη

Κάθε φορά που ένας χρήστης μέλος ακολουθεί τον σύνδεσμο ενός συγγράμματος προς τον διαδικτυακό τόπο του εκδότη, τότε στη νέα γραμμή που προστίθεται στον πίνακα JOURNAL_STATS, στο πεδίο EMAIL, καταχωρείται το πραγματικό του e-mail. Αυτό δηλαδή που χρησιμοποίησε για να δημιουργήσει το λογαριασμό του στο υποπεριβάλλον My-HEAL-Link.

2.6.1.5. Συνοπτικά οι τιμές των πεδίων SESSION και EMAIL για επισκέπτες και μέλη

Συνοψίζοντας τις τιμές που καταχωρούνται στα πεδία EMAIL και SESSION όταν οι χρήστες είναι επισκέπτες και μέλη, προκύπτει το ακόλουθο σχήμα.



Σχήμα 4.29. Οι τιμές των πεδίων EMAIL και SESSION για επισκέπτες και μέλη

2.6.2. Το πεδίο JOURNAL ως Item ID

Το item id αποτελεί τον κωδικό που προσδιορίζει μοναδικά τα αντικείμενα που περιλαμβάνονται στις συναλλαγές. Ο αλγόριθμος εξόρυξης θα πρέπει να το γνωρίζει ώστε να μπορεί να ξεχωρίζει τα αντικείμενα για να ανακαλύψει τις μεταξύ τους συσχετίσεις. Ως αντικείμενα στην περίπτωση των δεδομένων του HEAL-Link θεωρούνται τα ψηφιακά συγγράμματα. Η στήλη του πίνακα JOURNAL_STATS η οποία προσδιορίζει τα συγγράμματα που στοχοποιήθηκαν σε κάθε συναλλαγή, είναι η στήλη JOURNAL που αποθηκεύει τους τίτλους των συγγραμμάτων.

Όπως θα δούμε στη φάση μοντελοποίησης, εκτός από της στήλη JOURNAL του πίνακα JOURNAL_STATS που αποθηκεύει τους τίτλους των στοχοποιημένων συγγραμμάτων, θα χρειαστούμε και τους κωδικούς των συγγραμμάτων (στήλη J_ID του πίνακα JOURNAL) για την παραγωγή κανόνων συσχετίσεων με χρήση πληροφοριών ταξινομίας. Για να μπορέσουμε δηλαδή να μάθουμε σε ποιες θεματικές υποκατηγορίες και κατηγορίες ανήκουν τα στοχοποιημένα συγγράμματα, θα πρέπει να γνωρίζουμε τον κωδικό τους.

3. Επίλογος

Έχοντας κατανοήσει τα δεδομένα, την ιεραρχία με βάση την οποία γίνεται η αποθήκευση των ψηφιακών συγγραμμάτων και τον τρόπο λειτουργίας του συστήματος HEAL-Link, μπορούμε να προχωρήσουμε στην επόμενη φάση όπου θα προετοιμάσουμε τα δεδομένα προς εξόρυξη. Θα επιχειρήσουμε να ανιχνεύσουμε ημιτελή, θορυβώδη και ασυνεπή δεδομένα τα οποία θα τροποποιήσουμε κατάλληλα ώστε να είναι δυνατή η είσοδός τους στον αλγόριθμο εξόρυξης. Για την αντιμετώπιση των ατελειών στα δεδομένα ακολουθούμε διάφορους τρόπους, τους οποίους και περιγράφουμε αναλυτικά.

Κεφάλαιο 5. Προετοιμασία των δεδομένων

Τα δεδομένα όπως είναι διαθέσιμα από τις πραγματικές εφαρμογές στερούνται ποιότητας. Πολλές φορές οι χρήστες των βάσεων δεδομένων αναφέρουν λάθη, εμφάνιση ασυνήθιστων τιμών και ασυνέπειες στα αποθηκευμένα δεδομένα. Για το λόγο αυτόν είναι συνηθισμένο σε πραγματικές εφαρμογές τα υπό ανάλυση δεδομένα να είναι:

- *ημιτελή*: δηλαδή να παρουσιάζουν έλλειψη κάποιων τιμών γνωρισμάτων, έλλειψη ορισμένων γνωρισμάτων που ίσως έχουν ενδιαφέρον, ή να περιέχουν μόνο αθροιστικά δεδομένα,

- *θόρυβος*: περιέχουν λάθη ή outliers,

- *ασυνεπή*: περιέχουν αποκλίσεις στις κωδικοποιήσεις που χρησιμοποιούνται για να ταξινομήσουμε τα δεδομένα ή στα ονόματα που χρησιμοποιούνται για να αναφερθούμε στα ίδια δεδομένα.

Με βάση ένα σύνολο δεδομένων που στερείται ποιότητας τα αποτελέσματα της διαδικασίας εξόρυξης πληροφορίας αναπόφευκτα τείνουν να είναι ανακριβή και να μην παρουσιάζουν κάποιο ενδιαφέρον.

Η προ-επεξεργασία δεδομένων είναι ένα σημαντικό βήμα στη διαδικασία ανακάλυψης γνώσης. Οι τεχνικές προεπεξεργασίας δεδομένων που εφαρμόστηκαν πριν από το βήμα εξόρυξης πληροφορίας θα μπορούσαν να βοηθήσουν στη βελτίωση της ποιότητας των υπό ανάλυση δεδομένων και συνεπώς της ακρίβειας και της αποτελεσματικότητας των διαδικασιών που ακολουθούν της εξόρυξης πληροφορίας.

Υπάρχει ένας αριθμός από τεχνικές προ-επεξεργασίας δεδομένων που στοχεύει ουσιαστικά στη βελτίωση της γενικής ποιότητας της εξαγόμενης γνώσης. Οι πιο ευρύτατα χρησιμοποιούμενες τεχνικές μπορούν να συνοψιστούν στις ακόλουθες:

- *Καθαρισμός δεδομένων*, ο οποίος μπορεί να εφαρμοστεί για να αφαιρεθεί ο θόρυβος και να διορθωθούν τυχόν ασυνέπειες στα δεδομένα.

- *Μετασχηματισμός δεδομένων*. Μια κοινή τεχνική μετασχηματισμού είναι η κανονικοποίηση. Εφαρμόζεται προκειμένου να βελτιωθεί η ακρίβεια και η αποδοτικότητα των αλγορίθμων εξόρυξης που χρησιμοποιούν τις μετρήσεις απόστασης.

- *Μείωση δεδομένων*. Χρησιμοποιείται προκειμένου να μειωθεί το μέγεθος στοιχείων με συνάθροιση, εξαλείφοντας τα περιττά χαρακτηριστικά.

Στο κεφάλαιο αυτό εντοπίζουμε τα προβλήματα που υπάρχουν στα δεδομένα του HEAL-Link και τα καταγράφουμε μαζί με τους τρόπους αντιμετώπισής τους. Στόχος μας είναι να βελτιώσουμε την ποιότητα των υπό ανάλυση δεδομένων προτού τροφοδοτηθούν στον αλγόριθμο εξόρυξης ώστε τα αποτελέσματα της διαδικασίας ανακάλυψης γνώσης να είναι όσο το δυνατό πιο ακριβή.

1. Χρονοσφραγίδες σε μη επεξεργάσιμη μορφή

Οι χρονοσφραγίδες που είναι καταχωρημένες στο πεδίο TIME του πίνακα JOURNAL_STATS είναι αποθηκευμένες σε αλφαριθμητική μορφή (VARCHAR) γεγονός που τις καθιστά μη επεξεργάσιμες. Εάν θέλουμε να κάνουμε συγκρίσεις ή πράξεις μεταξύ δύο ή περισσότερων χρονοσφραγίδων δεν μπορούμε, γιατί η DB2 δεν τις αντιλαμβάνεται ως TIMESTAMPS αλλά ως απλές συμβολοσειρές. Για να λυθεί το πρόβλημα, θα πρέπει να βρούμε ένα τρόπο να μετατρέψουμε τις χρονοσφραγίδες σε μορφή επεξεργάσιμη από το σύστημα διαχείρισης της βάσης.

Στην προσπάθειά μας να μετατρέψουμε τις χρονοσφραγίδες σε μορφή κατάλληλη για εξόρυξη, θα πρέπει να λάβουμε υπόψιν το γεγονός ότι οι χρονοσφραγίδες που είναι αποθηκευμένες στον πίνακα JOURNAL_STATS δεν είναι όλες της ίδιας γλώσσας. Υπάρχουν χρονοσφραγίδες καταχωρημένες στα Αγγλικά και άλλες στα Ελληνικά. Επίσης, διαφέρουν ως προς τη μορφή, δηλαδή ως προς το υπόδειγμα (pattern) με βάση το οποίο δημιουργήθηκαν.

Υπάρχουν για παράδειγμα χρονοσφραγίδες της μορφής (Feb 18, 2009 3:36:52 PM) και άλλες της μορφής (Tue Jun 7 11:42:36 2005).

Όλα τα παραπάνω θα πρέπει να ληφθούν υπόψιν στην προσπάθεια μετατροπής των χρονοσφραγίδων.

1.1. TimestampConverter

Το πρόβλημα αντιμετωπίστηκε με την ανάπτυξη μίας εφαρμογής σε JAVA η οποία σχεδιάστηκε για να κάνει αυτήν ακριβώς τη δουλειά: Μετατροπή μη επεξεργάσιμων αλφαριθμητικών χρονοσφραγίδων, σε μορφή αναγνωρίσιμη και επεξεργάσιμη από τα συστήματα ανάλυσης δεδομένων. Η εφαρμογή ονομάζεται TimestampConverter.

1.1.1. Πώς λειτουργεί

Ο TimestampConverter δέχεται στην είσοδο χρονοσφραγίδες σε αλφαριθμητική μορφή, τις διασπά μία μία στα επιμέρους τμήματά τους και κατευθύνει τα τμήματα αυτά στην έξοδο. Η είσοδος και η έξοδος είναι δύο διαφορετικοί πίνακες της βάσης δεδομένων στην οποία έχει προηγουμένως συνδεθεί η εφαρμογή (υποστηρίζεται μόνο η DB2). Στην έξοδο δεν κατευθύνονται μόνο τα επιμέρους τμήματα της κάθε χρονοσφραγίδας αλλά και όλα τα υπόλοιπα δεδομένα του επιλεγμένου πηγαίου πίνακα.

Πιο αναλυτικά, τα επιμέρους τμήματα στα οποία διασπά ο TimestampConverter τις χρονοσφραγίδες είναι:

1. YEAR
2. MONTH
3. DAY
4. HOUR
5. MINUTE
6. SECOND
7. DISTANCE

Τα πρώτα έξι τμήματα είναι αυτά τα οποία συνθέτουν τη χρονοσφραγίδα. Το έβδομο τμήμα το υπολογίζει ο TimestampConverter. Είναι η χρονική απόσταση της χρονοσφραγίδας από την χρονική στιγμή 1/1/1970 00:00:00 UTC[1], δηλαδή τη χρονική στιγμή στην οποία ξεκίνησε να μετράει το ρολόι του UNIX, και υπολογίζεται σε milliseconds.

Ο λόγος για τον οποίο υπολογίζουμε την απόσταση αυτή είναι γιατί θα μας επιτρέψει να κάνουμε πολύ εύκολα υπολογισμούς και συγκρίσεις με τις χρονοσφραγίδες. Για παράδειγμα, θα μπορούμε πολύ εύκολα να υπολογίσουμε μέγιστες (νεότερες) και ελάχιστες (αρχαιότερες) χρονοσφραγίδες αλλά και χρονικές αποστάσεις.

1.1.2. Ορισμός παραμέτρου Locale

Ο TimestampConverter μας δίνει τη δυνατότητα να ρυθμίσουμε δύο βασικές παραμέτρους του αλγόριθμου μετατροπής.

Η πρώτη από αυτές είναι η παράμετρος Locale. Η παράμετρος Locale δίνει στον αλγόριθμο τη δυνατότητα να αναγνωρίζει χρονοσφραγίδες αποθηκευμένες σε διαφορετικές γλώσσες ορίζοντας έγκυρους κωδικούς γλωσσών κατά ISO, όπως για παράδειγμα "el" για τα Ελληνικά και "en" για τα Αγγλικά. Για κάθε κωδικό που ορίζουμε στην παράμετρο αυτή, ο αλγόριθμος θα μπορεί να μετατρέψει χρονοσφραγίδες που δημιουργήθηκαν στη γλώσσα που αναπαριστά ο συγκεκριμένος κωδικός.

Ο χρήστης μπορεί να προσθέσει ή να αφαιρέσει κωδικούς από τη λίστα επιλογών. Οι δύο προεπιλεγμένοι κωδικοί που έχει το πρόγραμμα κατά την εκκίνησή του είναι οι "el" για την Ελληνική γλώσσα και "en" για την Αγγλική.

[1]Standard base time known as "the epoch", namely January 1, 1970, 00:00:00 GMT (Unix time).

Με αυτή τη δυνατότητα του προγράμματος TimestampConverter λύνουμε το πρόβλημα της μετατροπής χρονοσφραγίδων που είναι αποθηκευμένες σε διαφορετικές γλώσσες.

1.1.3. Ορισμός παραμέτρου Pattern

Η δεύτερη παράμετρος, Pattern (υπόδειγμα), δίνει στον αλγόριθμο τη δυνατότητα να αναγνωρίζει χρονοσφραγίδες τυπωμένες σε διαφορετικές διατάξεις (formats). Κάθε χρονοσφραγίδα, όταν δημιουργείται αποκτά μία συγκεκριμένη διάταξη σύμφωνα με το υπόδειγμα που χρησιμοποιήθηκε κατά τη δημιουργία της. Λέγοντας pattern, εννοούμε μία συμβολοσειρά που αποτελείται από κατάλληλο συνδυασμό των ακόλουθων χαρακτήρων (ο πίνακας προέρχεται από το API της Java: http://download.oracle.com/javase/6/docs/api/java/text/SimpleDateFormat.html)

Πίνακας 5.1. Συντακτικό για τη δημιουργία patterns

Symbol	Meaning	Presentation	Example
G	Era designator	Text	AD
y	Year	Year	1996; 96
M	Month in year	Month	July; Jul; 07
w	Week in year	Number	27
W	Week in month	Number	2
D	Day in year	Number	189
d	Day in month	Number	10
F	Day of week in month	Number	2
E	Day in week	Text	Tuesday; Tue
a	Am/pm marker	Text	PM
H	Hour in day (0-23)	Number	0
k	Hour in day (1-24)	Number	24
K	Hour in am/pm (0-11)	Number	0
h	Hour in am/pm (1-12)	Number	12
m	Minute in hour	Number	30
s	Second in minute	Number	55
S	Millisecond	Number	978
z	Time zone	General time zone	Pacific Standard Time; PST; GMT-08:00
Z	Time zone	RFC 822 time zone	-0800

Αν για παράδειγμα υπάρχει στη βάση δεδομένων η χρονοσφραγίδα

2009.06.30 AD at 08:29:36 PDT

και θέλουμε να τη μετατρέψουμε, θα πρέπει να πούμε στον TimestampConverter να λάβει υπόψιν του κατά τη μετατροπή το ακόλουθο υπόδειγμα:

yyyy.MM.dd G 'at' hh:mm:ss z

Το πρόγραμμα έχει τρία προεπιλεγμένα υποδείγματα τα οποία είναι έτοιμα προς χρήση. Ο χρήστης μπορεί να προσθέσει και άλλα κάνοντας κλικ στο κουμπί με τον μπλε σταυρό:

Πίνακας 5.2. Προεπιλεγμένα υποδείγματα του TimestampConverter

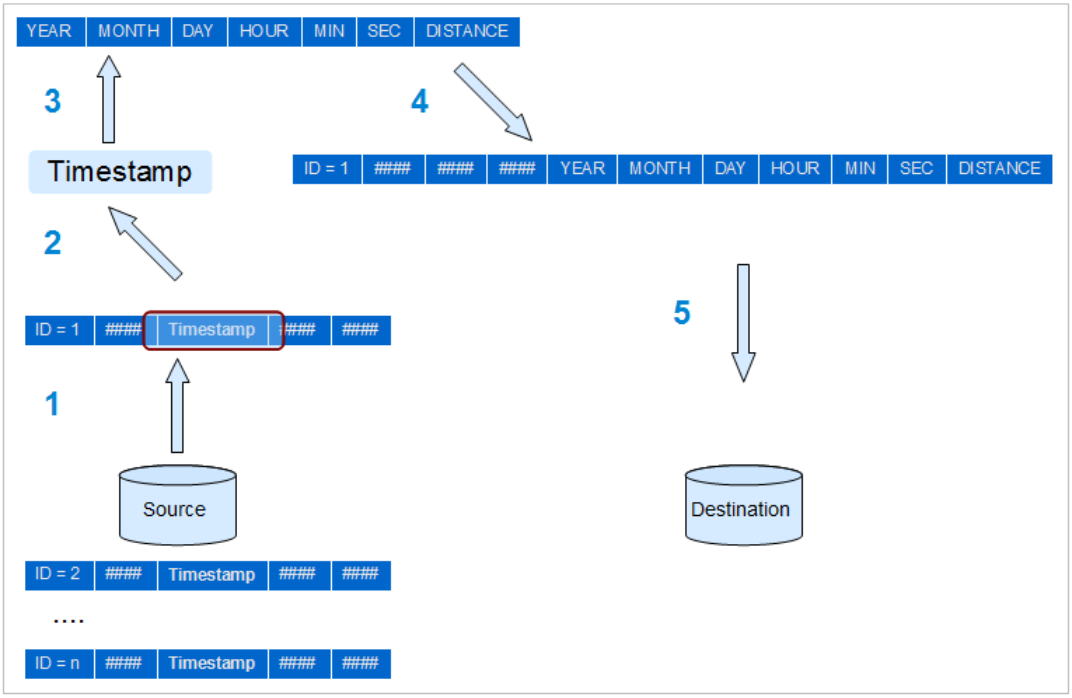
Pattern	Output
MMM d, yyyy K:m:s a	Aug 15, 2010 08:10:33 am
d MMM yyyy K:m:s a	8 Dec 2009 10:40:20 pm

Pattern	Output
EEE MMM d k:m:s yyyy	Tue Dec 19 18:20:14 2007

Με τη δυνατότητα του TimestampConverter να διαχειρίζεται διαφορετικά patterns λύνουμε και το δεύτερο πρόβλημα, αυτό της ύπαρξης χρονοσφραγίδων που δημιουργήθηκαν με διαφορετικά υποδείγματα.

1.1.4. Ο αλγόριθμος μαζικής μετατροπής

Στο ακόλουθο σχήμα φαίνεται η ροή των δεδομένων και τα βήματα του αλγόριθμου κατά τη μαζική μετατροπή.



Σχήμα 5.1. Ροή δεδομένων και βήματα του αλγόριθμου μαζικής μετατροπής

Ο αλγόριθμος μετατροπής εκτελεί κατά σειρά τα ακόλουθα βήματα:

Για κάθε εγγραφή του πηγαίου πίνακα:

1. Ανάκτηση της πρώτης κατά σειρά εγγραφής από τον πηγαίο πίνακα (όχι μόνο της χρονοσφραγίδας αλλά ολόκληρης της εγγραφής).

2. Απομόνωση της χρονοσφραγίδας από την ανακτημένη εγγραφή (απομονώνεται η χρονοσφραγίδα για να διασπαστεί στα επιμέρους τμήματά της).

3. Διάσπαση της χρονοσφραγίδας στα επιμέρους τμήματά της και υπολογισμός της χρονικής απόστασης από την 1/1/1970 00:00:00 UTC.

4. Σύνθεση της νέας εγγραφής που θα αποθηκευτεί στον πίνακα προορισμού. Η νέα εγγραφή αποτελείται από τα ίδια ακριβώς πεδία της παλιάς με την εξής διαφορά: Το πεδίο που περιλαμβάνει τη χρονοσφραγίδα αντικαθίσταται από τα εφτά νέα πεδία που παρήγαγε ο αλγόριθμος (year, month, day, hour, minute, second, distance).

5. Καταχώρηση της νέας εγγραφής στον πίνακα προορισμού.

1.1.5. Οι περιορισμοί του αλγόριθμου

Από τα παραπάνω βήματα προκύπτουν οι ακόλουθοι περιορισμοί για τους οποίους θα πρέπει να φροντίσει ο χρήστης διότι δεν ικανοποιούνται αυτόματα από το πρόγραμμα:

1. Ο πίνακας ο οποίος θα οριστεί ως πηγή χρονοσφραγίδων θα πρέπει να έχει ένα πεδίο με όνομα "ID". Το πεδίο αυτό θα πρέπει να περιέχει μία μοναδική και αδιάσπαστη ακολουθία τιμών κατά αύξουσα σειρά ξεκινώντας από την τιμή 1. Για παράδειγμα, η πρώτη εγγραφή του πίνακα θα έχει στο πεδίο ID τιμή 1, η δεύτερη τιμή 2, η νιοστή τιμή ν κλπ. Ένας καλός τρόπος για να γίνει αυτό στην DB2 είναι να οριστεί η στήλη ID ως Identity Column.

   Καλό θα ήταν να προστεθεί στο πεδίο ID και ένα INDEX το οποίο θα συμβάλλει κατά πολύ στη βελτίωση των επιδόσεων.

2. Ο πίνακας προορισμού θα πρέπει να έχει την ίδια ακριβώς δομή με τον πηγαίο πίνακα με την εξής διαφορά: Η στήλη που περιέχει τις χρονοσφραγίδες θα πρέπει να αντικατασταθεί από τα ακόλουθα πεδία τα οποία θα φιλοξενήσουν τα τμήματα της διασπασμένης χρονοσφραγίδας:

   - YEAR INTEGER NOT NULL
   - MONTH INTEGER NOT NULL
   - DAY INTEGER NOT NULL
   - HOUR INTEGER NOT NULL
   - MINUTE INTEGER NOT NULL
   - SECOND INTEGER NOT NULL
   - DISTANCE BIGINT NOT NULL

3. Ο πίνακας προορισμού θα πρέπει να είναι τελείως άδειος πριν την εκκίνηση του αλγόριθμου μαζικής μετατροπής. Εάν δεν είναι, ενδέχεται να προκύψουν παραβιάσεις περιορισμών κύριου κλειδιού.

1.2. Προετοιμασία για την λειτουργία του TimestampConverter

Με βάση τους περιορισμούς που θέτει ο TimestampConverter, θα πρέπει να προετοιμάσουμε κατάλληλα τη βάση δεδομένων. Συγκεκριμένα, θα πρέπει να εκτελέσουμε τις ακόλουθες ενέργειες:

1.2.1. Δημιουργία του πηγαίου πίνακα JOURNAL_STATS_WITH_ID

Ο πίνακας JOURNAL_STATS_WITH_ID θα αποτελέσει τον πηγαίο πίνακα, το σημείο δηλαδή από το οποίο το πρόγραμμα θα αντλεί τις χρονοσφραγίδες.

Σύμφωνα με τον πρώτο περιορισμό του TimestampConverter, ο πηγαίος πίνακας θα πρέπει να έχει ένα πεδίο με όνομα "ID" το οποίο θα περιέχει μία μοναδική και αδιάσπαστη ακολουθία τιμών κατά αύξουσα σειρά ξεκινώντας από την τιμή 1.

Ο κώδικας SQL για τη δημιουργία αυτού του πίνακα είναι:

```
CREATE TABLE JOURNAL_STATS_WITH_ID (
  ID BIGINT NOT NULL GENERATED BY DEFAULT AS IDENTITY ( START WITH 1 INCREMENT
   BY 1 MINVALUE 1 MAXVALUE 9223372036854775807 NO CYCLE CACHE 20),
  EMAIL VARCHAR(80) NOT NULL,
  IP VARCHAR(25) NOT NULL,
  TIME VARCHAR(50) NOT NULL,
  TYPE VARCHAR(10) NOT NULL,
  JOURNAL VARCHAR(255) NOT NULL,
  SESSION VARCHAR(255)
 )

CREATE UNIQUE INDEX JOURNAL_STATS_WITH_ID__IDX
 ON JOURNAL_STATS_WITH_ID
 (ID  ASC) PCTFREE 0
ALLOW REVERSE SCANS;
```

```
ALTER TABLE JOURNAL_STATS_WITH_ID ADD CONSTRAINT SQL100801195800920 PRIMARY KEY
 (IP,
  TIME,
  TYPE,
  JOURNAL);
```

1.2.2. Δημιουργία του πίνακα προορισμού JOURNAL_STATS_WITH_ID_WCT

Ο πίνακας JOURNAL_STATS_WITH_ID_WCT (With Converted Timestamps) θα αποτελέσει τον πίνακα προορισμού. Το σημείο δηλαδή στο οποίο θα κατευθυνθούν όλα τα δεδομένα του πηγαίου πίνακα με τις χρονοσφραγίδες διασπασμένες στα επιμέρους τμήματά τους. Έχει την ίδια δομή με αυτή του πηγαίου πίνακα με μία μόνο διαφορά. Το πεδίο TIME αντικαθίσταται από τα εφτά πεδία που θα φιλοξενήσουν τα τμήματα των χρονοσφραγίδων και την χρονική απόστασή τους από την 1/1/1970 1/1/1970 00:00:00 UTC.

Ο κώδικας για τη δημιουργία του είναι:

```
CREATE TABLE JOURNAL_STATS_WITH_ID_WCT (
  ID BIGINT NOT NULL,
  EMAIL VARCHAR(80) NOT NULL,
  IP VARCHAR(25) NOT NULL,
  TYPE VARCHAR(10) NOT NULL,
  JOURNAL VARCHAR(255) NOT NULL,
  SESSION VARCHAR(255),
  YEAR INTEGER NOT NULL,
  MONTH INTEGER NOT NULL,
  DAY INTEGER NOT NULL,
  HOUR INTEGER NOT NULL,
  MINUTE INTEGER NOT NULL,
  SECOND INTEGER NOT NULL,
  DISTANCE BIGINT NOT NULL
 )

ALTER TABLE JOURNAL_STATS_WITH_ID_WCT ADD CONSTRAINT SQL100801195800920 PRIMARY KEY
 (IP,
  TYPE,
  JOURNAL,
  YEAR,
  MONTH,
  DAY,
  HOUR,
  MINUTE,
  SECOND);
```

1.3. Μετατροπή των χρονοσφραγίδων

Ακολουθούν τα βήματα για την έναρξη της διαδικασίας μετατροπής των χρονοσφραγίδων.

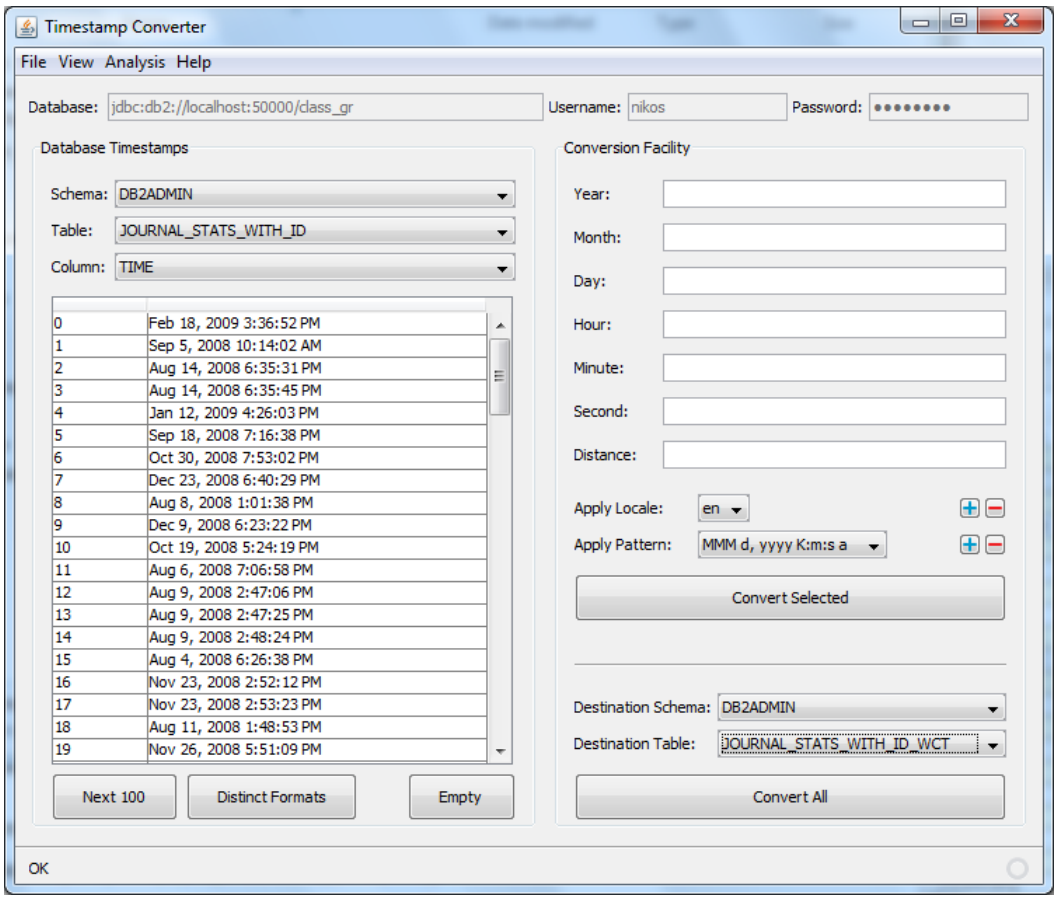
1. Εκκινούμε την εφαρμογή TimestampConverter και συνδεόμαστε στη βάση δεδομένων.

2. Επιλέγουμε κατά σειρά το σχήμα, τον πίνακα και τη στήλη που φιλοξενούν τις χρονοσφραγίδες (JOURNAL_STATS_WITH_ID στήλη TIME).

3. Αφού ελέγξουμε και βεβαιωθούμε για την ορθότητα των παραμέτρων μετατροπής, στη συνέχεια ορίζουμε το σχήμα και τον πίνακα προορισμού (JOURNAL_STATS_WITH_ID_WCT).

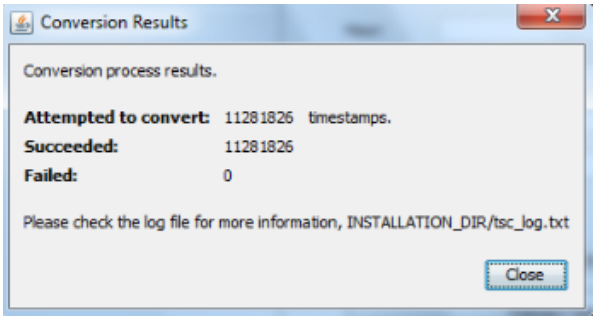
4. Όλα είναι έτοιμα για να ξεκινήσει η μετατροπή. Κάνουμε κλικ στο κουμπί "Convert All" και το πρόγραμμα ξεκινάει να μετατρέπει τις χρονοσφραγίδες.

Στο ακόλουθο σχήμα φαίνονται οι επιλογές που κάναμε πριν την έναρξη της διαδικασίας μετατροπής.



Σχήμα 5.2. Πριν την έναρξη της διαδικασίας μετατροπής

Για να μετατρέψει ο TimestampConverter όλες τις χρονοσφραγίδες (11.281.826) λειτουργούσε συνεχόμενα επί τρία εικοσιτετράωρα. Τελικά, ολοκλήρωσε με επιτυχία επιστρέφοντας το ακόλουθο αποτέλεσμα.



Σχήμα 5.3. Μετά την ολοκλήρωση της μετατροπής

Ένα δείγμα των δεδομένων του πίνακα JOURNAL_STATS_WITH_ID_WCT φαίνεται στο ακόλουθο σχήμα:

	ID	EMAIL	IP	TYPE	JOURNAL	SESSION	YEAR	MONTH	DAY	HOUR	MINUTE	SECOND	DISTANCE
1	11265...	guest	155.207.8...	alpha	American Journal o...	036091A8...	2009	2	24	16	26	29	1235485589000
2	11265...	guest	83.235.17...	alpha	American Journal o...	5659FE64...	2009	2	24	16	43	45	1235486625000
3	11265...	guest	155.207.8...	alpha	American Journal o...	036091A8...	2009	2	24	16	31	34	1235485894000
4	11265...	guest	147.102.2...	alpha	Biological Reviews ...	628CAEEF...	2009	2	24	16	25	13	1235485513000
5	11265...	guest	147.102.2...	alpha	Breast Cancer Online	628CAEEF...	2009	2	24	16	35	41	1235486141000
6	11265...	guest	147.102.2...	alpha	Bulletin of the Lon...	628CAEEF...	2009	2	24	16	29	19	1235485759000
7	11265...	guest	147.102.3...	alpha	Chronicle of Highe...	8B75E339...	2009	2	24	16	19	11	1235485151000
8	11265...	guest	147.102.3...	alpha	Chronicle of Highe...	8B75E339...	2009	2	24	16	24	12	1235485452000
9	11265...	guest	139.91.19...	alpha	Clinical Physics an...	6D8CD20...	2009	2	24	16	39	55	1235486395000
10	11265...	guest	195.251.3...	alpha	Common Market L...	C81C0791...	2009	2	24	16	26	45	1235485605000

Σχήμα 5.4. Δείγμα δεδομένων του πίνακα JOURNAL_STATS_WITH_ID_WCT

2. Πεδίο SESSION

Το πεδίο SESSION χρήζει ιδιαίτερης προσοχής στην φάση της προετοιμασίας που βρισκόμαστε. Θα είναι σημείο κλειδί για τον αλγόριθμο παραγωγής κανόνων συσχετίσεων αφού αποτελεί το transaction id των συναλλαγών που εκτελούνται από επισκέπτες. Συνεπώς, θα πρέπει να το εξετάσουμε σε βάθος και να βεβαιωθούμε ότι τα δεδομένα τα οποία φέρει είναι έτοιμα να τροφοδοτηθούν στον αλγόριθμο εξόρυξης.

Εκτός από τους μοναδικούς κωδικούς των συνεδριών και το αλφαριθμητικό 'ok', διαπιστώσαμε ότι καταχωρείται στο πεδίο SESSION και το NULL.

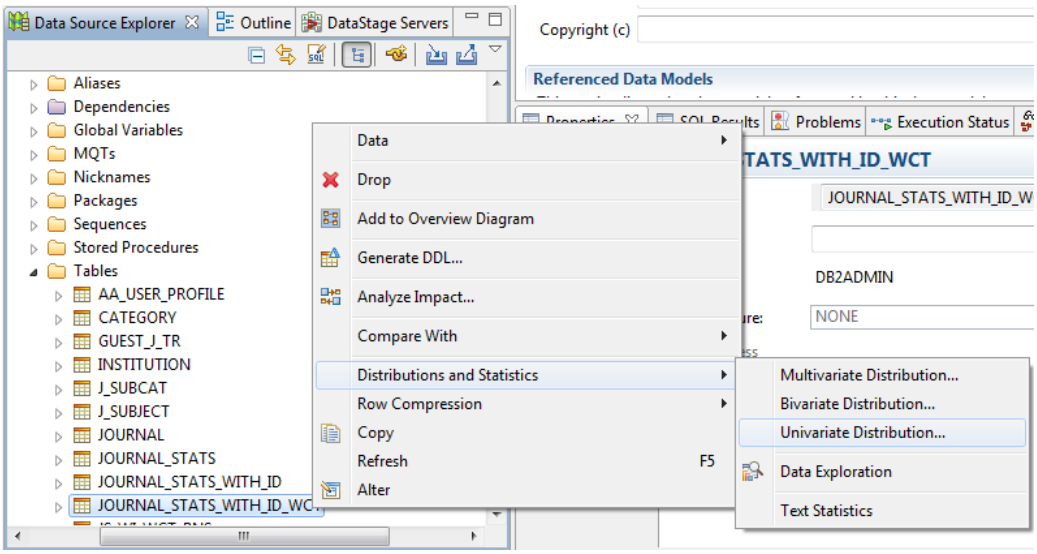
Στην προσπάθειά μας να εξερευνήσουμε τις διαφορετικές τιμές του πεδίου SESSION και την κατανομή τους, θα χρησιμοποιήσουμε το Design Studio και τη λειτουργία "Univariate Distribution". Η λειτουργία αυτή αναλύει τα δεδομένα του πίνακα στον οποίο εφαρμόζεται και παράγει ένα σύνολο γραφημάτων που παρουσιάζουν την κατανομή των τιμών σε κάθε στήλη του πίνακα.

Στη προκειμένη περίπτωση, θα εκτελέσουμε τη λειτουργία "Univariate Distribution" στον πίνακα JOURNAL_STATS_WITH_ID_WCT για να μάθουμε την κατανομή όλων των τιμών του πεδίου SESSION ανά έτος.

2.1. Πλήθος εγγραφών με session = null ανά έτος

Για να ανακαλύψουμε το πλήθος των εγγραφών που έχουν NULL στο πεδίο SESSION, εκτελούμε τα ακόλουθα βήματα:

1. Στον “Data Source Explorer”, δεξί κλικ στον πίνακα JOURNAL_STATS_WITH_ID_WCT > Distributions and Statistics > Univariate Distribution...



Σχήμα 5.5. Λειτουργία Univariate Distribution

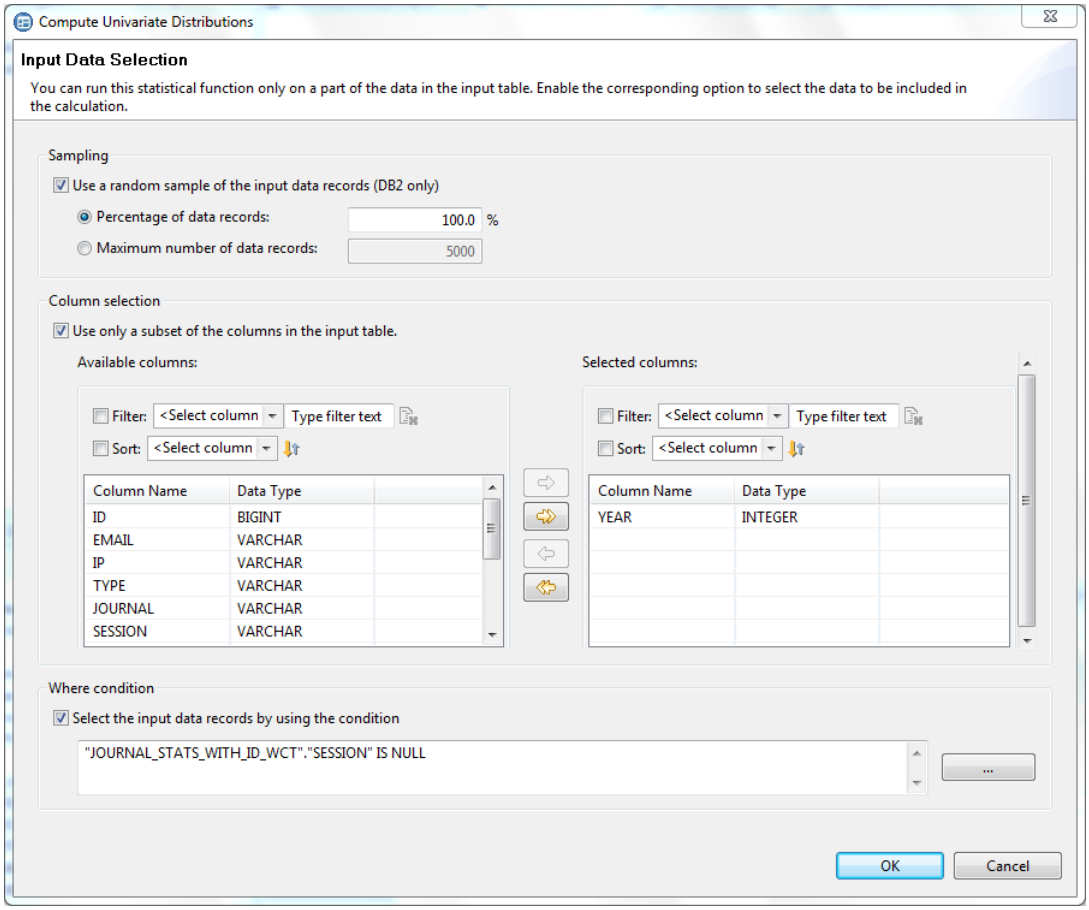
2. Στο “Input Data Selection”, επιλέγουμε “Percentage of data records” και εισάγουμε ποσοστό 100% (επιθυμούμε εφαρμογή σε όλα τα δεδομένα).

3. Στο “Column selection”, ενεργοποιούμε την επιλογή “Use only a subset of the columns in the input table” και αφαιρούμε από τη λίστα “Selected columns:” όλες τις στήλες εκτός από το YEAR (θέλουμε να μάθουμε την κατανομή των εγγραφών στο πεδίο YEAR).

4. Στο “Where condition”, ενεργοποιούμε την επιλογή “Select the input data records by using the condition” και εισάγουμε τη συνθήκη:

```
"JOURNAL_STATS_WITH_ID_WCT"."SESSION" IS NULL
```

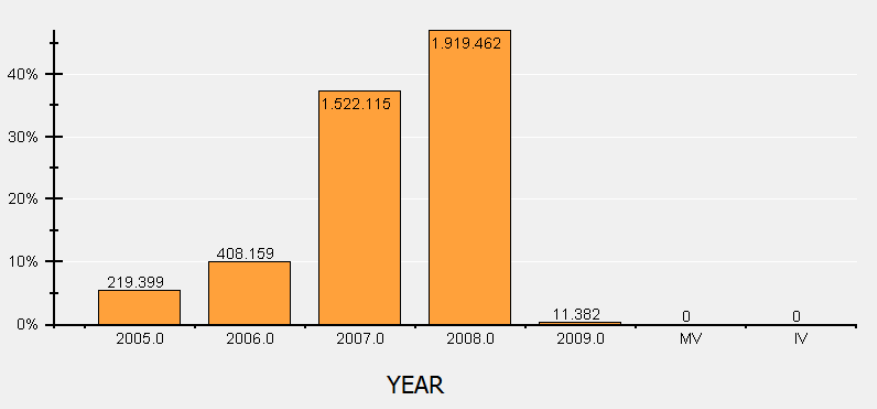
   Επιθυμούμε εφαρμογή μόνο σε εκείνες τις εγγραφές που έχουν NULL στο πεδίο SESSION. Μπορούμε να διαμορφώσουμε τη συνθήκη πατώντας στο κουμπί με τα αποσιωπητικά (...).



Σχήμα 5.6. Λίγο πριν την εκτέλεση της λειτουργίας Univariate Distribution

5. Πατάμε το κουμπί OK.

Η συνάρτηση εκτελείται και επιστρέφει το ακόλουθο γράφημα. Στο γράφημα βλέπουμε πόσες από τις εγγραφές που έχουν NULL στο πεδίο SESSION καταχωρήθηκαν σε κάθε έτος.



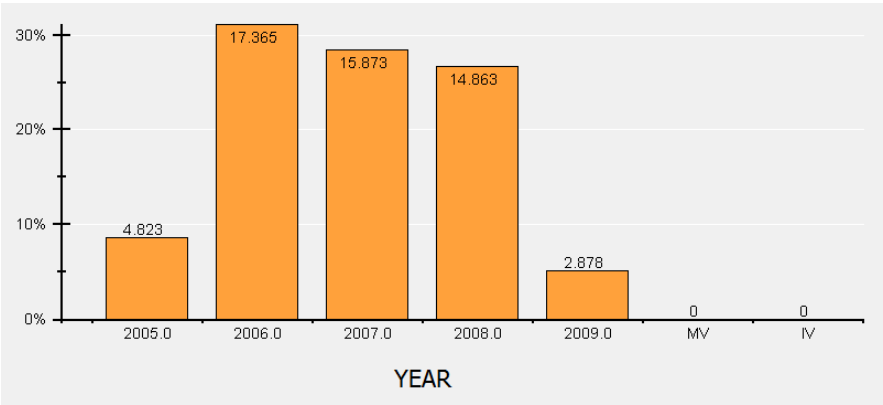
Σχήμα 5.7. Πλήθος εγγραφών με session = null ανά έτος

2.2. Πλήθος εγγραφών με session = 'ok' ανά έτος

Τα βήματα είναι ακριβώς τα ίδια. Το μόνο που αλλάζει είναι η συνθήκη στο βήμα 4:

```
"JOURNAL_STATS_WITH_ID_WCT"."SESSION" = 'ok'
```

Το γράφημα μας δείχνει πώς κατανέμονται ανά έτος οι εγγραφές που έχουν στο πεδίο SESSION την τιμή 'ok'. Δηλαδή όλες εκείνες οι εγγραφές που καταχωρήθηκαν από εγγεγραμμένα μέλη.



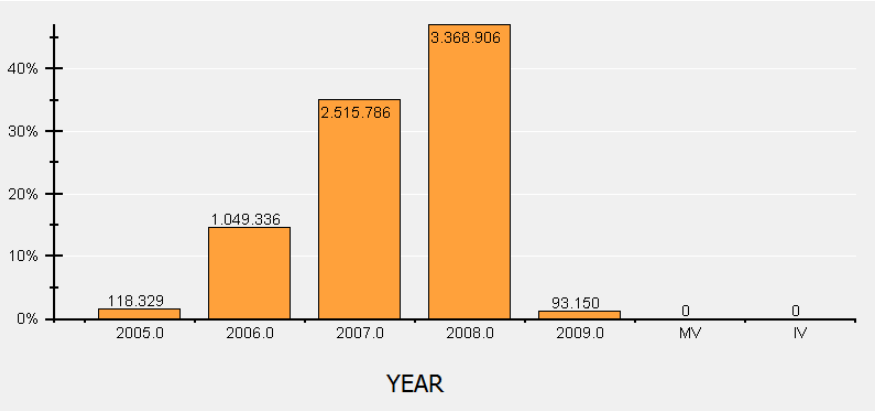
Σχήμα 5.8. Πλήθος εγγραφών με session = 'ok' ανά έτος

2.3. Πλήθος εγγραφών με session = session_id ανά έτος

Και πάλι τα βήματα είναι τα ίδια. Αλλάζει μόνο είναι η συνθήκη στο βήμα 4:

```
"JOURNAL_STATS_WITH_ID_WCT"."SESSION" IS NOT NULL AND "JOURNAL_STATS_WITH_ID_WCT"."SESSION" NOT LIKE 'ok'
```

Το γράφημα μας δείχνει πώς κατανέμονται ανά έτος οι εγγραφές που έχουν στο πεδίο SESSION μοναδικό session_id. Δηλαδή όλες εκείνες οι εγγραφές που καταχωρήθηκαν από επισκέπτες.



Σχήμα 5.9. Πλήθος εγγραφών με session = session_id ανά έτος

2.4. Συγκεντρωτικά τα αποτελέσματα

Ο πίνακας 5.3 συγκεντρώνει τα παραπάνω αποτελέσματα.

Πίνακας 5.3. Πλήθος εγγραφών με βάση το session ανά έτος (JOURNAL_STATS_WITH_ID_WCT)

Έτος	session = null	session = 'ok'	session = session_id	session != null	Σύνολο
2005	219.399	4.823	118.329	123.152	342.551
2006	408.159	17.365	1.049.336	1.066.701	1.474.860
2007	1.522.115	15.873	2.515.786	2.531.659	4.053.774
2008	1.919.462	14.863	3.368.906	3.383.769	5.303.231
2009	11.382	2.878	93.150	96.028	107.410
Σύνολο	**4.080.517**	**55.802**	**7.145.507**	**7.201.309**	**11.281.826**

2.5. Αφαίρεση των εγγραφών που έχουν NULL στο πεδίο SESSION

Με βάση τον διαχωρισμό των δεδομένων συναλλαγών σε δεδομένα επισκεπτών και δεδομένα εγγεγραμμένων μελών, όλες εκείνες τις εγγραφές του πίνακα JOURNAL_STATS_WITH_ID_WCT που έχουν NULL στο πεδίο SESSION θα μπορούσαμε να τις αφαιρέσουμε καθώς δεν είναι ξεκάθαρη η προέλευσή τους (προέρχονται από επισκέπτες ή από εγγεγραμμένα μέλη;). Ο διαχωρισμός των δεδομένων είναι σαφής:

- Εγγραφές με SESSION = session_id και e-mail = 'guest' προέρχονται από επισκέπτες.
- Εγγραφές με SESSION = 'ok' και πραγματικό e-mail προέρχονται από εγγεγραμμένα μέλη.

Όλες τις άλλες περιπτώσεις (SESSION = NULL) θα επιχειρήσουμε να τις αφαιρέσουμε ώστε να είναι ξεκάθαρη η εικόνα των δεδομένων πριν την ανάλυσή τους. Βέβαια, θα μπορούσαμε να τις κρατήσουμε και να τις συμπεριλάβουμε στην ανάλυση, κάτι τέτοιο όμως θα μας ανάγκαζε να παράγουμε πολλά διαφορετικά μοντέλα συσχετίσεων (ένα για κάθε διαφορετική ομάδα δεδομένων συναλλαγών) γεγονός το οποίο θα αύξανε σε μεγάλο βαθμό την πολυπλοκότητα του έργου χωρίς να παρέχει ιδιαίτερα πλεονεκτήματα.

Δημιουργούμε ένα νέο πίνακα με όνομα JS_WI_WCT_RNS (Journal-Stats-With-ID-With-Converted-Timestamps-Removed-Null-Sessions) και τον τροφοδοτούμε με όλα τα δεδομένα του πίνακα JOURNAL_STATS_WITH_ID_WCT, εξαιρώντας τις εγγραφές που έχουν null στο πεδίο session.

Ακολουθεί ο κώδικας τροφοδοσίας του πίνακα JS_WI_WCT_RNS:

```
insert
into db2admin.JS_WI_WCT_RNS (
select *
```

```
from db2admin.JOURNAL_STATS_WITH_ID_WCT
where session is not null
);
```

Τα νέα δεδομένα διαμορφώνονται ως εξής:

Πίνακας 5.4. Πλήθος εγγραφών με βάση το session ανά έτος (JS_WI_WCT_RNS)

Έτος	session = null	session = 'ok'	session = session_id	Σύνολο
2005	0	4.823	118.329	123.152
2006	0	17.365	1.049.336	1.066.701
2007	0	15.873	2.515.786	2.531.659
2008	0	14.863	3.368.906	3.383.769
2009	0	2.878	93.150	96.028
Σύνολο	**0**	**55.802**	**7.145.507**	**7.201.309**

Στη συνέχεια θα μετρήσουμε το πλήθος των μοναδικών συναλλαγών που είναι καταχωρημένες στον πίνακα JS_WI_WCT_RNS. Για να γίνει αυτό θα πρέπει να εκτελέσουμε δύο διαφορετικά ερωτήματα SQL. Το πρώτο για να μετρήσουμε τις μοναδικές συναλλαγές που εκτελέστηκαν από τα εγγεγραμμένα μέλη και το δεύτερο για την καταμέτρηση αυτών που εκτελέστηκαν από επισκέπτες.

Ο λόγος για τον οποίο εκτελούμε δύο ερωτήματα είναι γιατί τα transaction id των δύο διαφορετικών τύπων συναλλαγών διαφέρουν. Για τους μεν εγγεγραμμένους χρήστες ως transaction id θεωρούμε το συνδυασμό των πεδίων EMAIL, YEAR, MONTH και DAY, για τους δε επισκέπτες το transaction id είναι το ίδιο με το session id. Συνεπώς, τα ερωτήματα διαμορφώνονται ως εξής:

Για τα εγγεγραμμένα μέλη, ομαδοποιούμε τις εγγραφές που έχουν 'ok' στο πεδίο SESSION με βάση τα πεδία EMAIL, YEAR, MONTH και DAY. Στη συνέχεια μετράμε το πλήθος των ομάδων (κάθε group και μία ξεχωριστή συναλλαγή).

```
select count(*)
from (
select EMAIL, YEAR, MONTH, DAY
from db2admin.JS_WI_WCT_RNS
where session like 'ok'
group by EMAIL, YEAR, MONTH, DAY
) groups;
```

Για τους επισκέπτες, εξαιρούμε τις εγγραφές που έχουν 'ok' στο πεδίο SESSION και μετράμε το πλήθος των μοναδικών session id.

```
select count(distinct session)
from db2admin.JS_WI_WCT_RNS
where session not like 'ok'
```

Ο ακόλουθος πίνακας μας δείχνει το πλήθος των μοναδικών συναλλαγών που είναι καταχωρημένες στον πίνακα JS_WI_WCT_RNS.

Πίνακας 5.5. Πλήθος μοναδικών συναλλαγών στον πίνακα JS_WI_WCT_RNS

Συναλλαγές	Πλήθος
Μελών	17.386
Επισκεπτών	5.834.658

3. Εγγραφές που έχουν κενό (' ') στο πεδίο EMAIL

Στο πεδίο EMAIL του πίνακα JS_WI_WCT_RNS, εκτός από πραγματικές ηλεκτρονικές διευθύνσεις που καταχωρούνται από τα εγγεγραμμένα μέλη και το αλφαριθμητικό 'guest' που καταχωρείται από τους επισκέπτες, εντοπίστηκε μία ακόμη μοναδική τιμή. Το αλφαριθμητικό κενό (' ').

Η παραπάνω διαπίστωση προκύπτει από την εκτέλεση του SQL ερωτήματος:

```
select distinct email
from db2admin.JS_WI_WCT_RNS
where email not like '%@%';
```

Δημιουργείται σε αυτό το σημείο το ερώτημα:

- Πότε καταχωρείται η τιμή αυτή και γιατί;

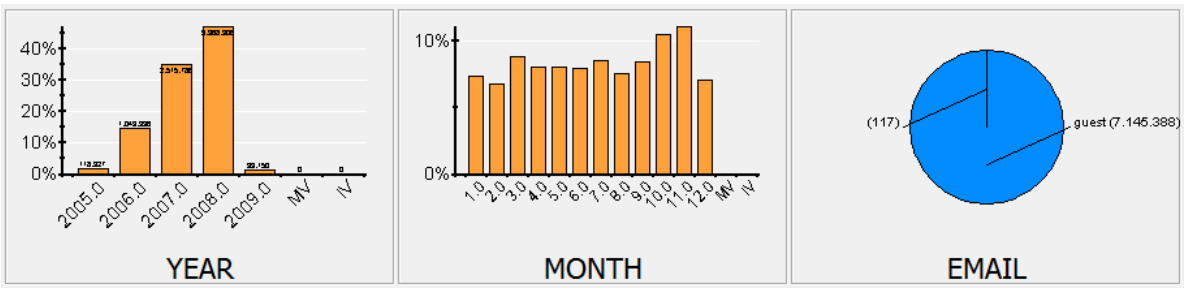
Για να μπορέσουμε να απαντήσουμε, θα χρησιμοποιήσουμε το Design Studio και τη λειτουργία “Univariate Distribution”. Στη συγκεκριμένη περίπτωση, θέλουμε να εκτελεστεί η λειτουργία σε εκείνες τις εγγραφές που δεν έχουν πραγματικό email καταχωρημένο στο πεδίο EMAIL. Ο λόγος για τον οποίο θέτουμε αυτόν τον περιορισμό είναι για να εξαιρέσουμε από την ανάλυση τους χρήστες που είναι εγγεγραμμένα μέλη του υποπεριβάλλοντος My-HEAL-Link. Θέλουμε να μελετήσουμε την κατανομή μόνο των τιμών 'guest και ' '. Ο περιορισμός αυτός υλοποιείται στο τρίτο από τα τέσσερα ακόλουθα βήματα:

1. Στον “Data Source Explorer”, δεξί κλικ στον πίνακα JS_WI_WCT_RNS > Distributions and Statistics > Univariate Distribution...
2. Στο “Input Data Selection”, επιλέγουμε “Percentage of data records” και εισάγουμε ποσοστό 100% (επιθυμούμε εφαρμογή σε όλα τα δεδομένα).
3. Στο “Where condition”, τσεκάρουμε την επιλογή “Select the input data records by using the condition” και εισάγουμε τη συνθήκη:

   ```
   "JS_WI_WCT_RNS"."EMAIL" NOT LIKE '%@%'
   ```

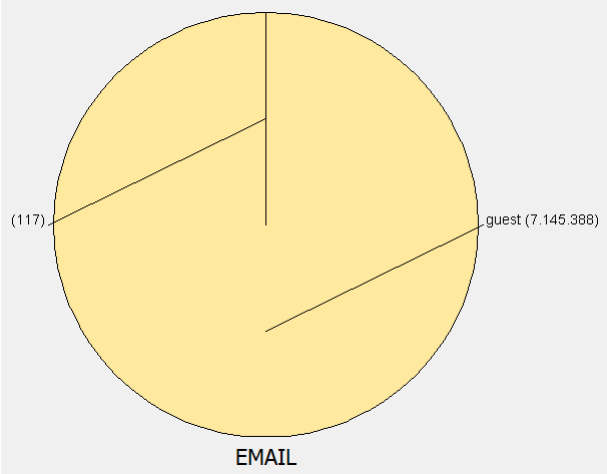
   Μπορούμε να διαμορφώσουμε τη συνθήκη πατώντας στο κουμπί με την έλλειψη (...).
4. Πατάμε OK.

Μόλις ολοκληρωθεί η διαδικασία εμφανίζονται, μεταξύ άλλων, και τα ακόλουθα γραφήματα.



Σχήμα 5.10. “Univariate Distribution” στον πίνακα JS_WI_WCT_RNS

Το τρίτο κατά σειρά γράφημα είναι αρκετά διαφωτιστικό και παρουσιάζεται στη συνέχεια.



Σχήμα 5.11. Κατανομή τιμών 'guest' και κενό (' ') στο πεδίο EMAIL

Η πρώτη πληροφορία την οποία αντλούμε από το γράφημα είναι ότι η τιμή 'guest' είναι καταχωρημένη στο πεδίο EMAIL 7.145.388 φορές ενώ η τιμή ' ' μόλις 117. Συνεπώς, μπορούμε με ασφάλεια να συμπεράνουμε, απαντώντας στο ερώτημα που τέθηκε παραπάνω, ότι η τιμή 'guest' είναι αυτή η οποία καταχωρείται στο πεδίο EMAIL όταν ένας επισκέπτης στοχοποιεί ένα σύγγραμμα και όχι το κενό, αφού η πρώτη συναντάται σχεδόν στο 100% των εγγραφών.

Τις 117 εγγραφές που καταχωρούν το κενό στο πεδίο EMAIL θα μπορούσαμε να τις χαρακτηρίσουμε ως έκτοπα και θα τις αφαιρέσουμε από τον πίνακα JS_WI_WCT_RNS.

Δημιουργούμε ένα νέο πίνακα με όνομα JS_WI_WCT_RNS_V2 (version 2) και αντιγράφουμε σε αυτόν όλα τα περιεχόμενα του πίνακα JS_WI_WCT_RNS εξαιρώντας τις εγγραφές που έχουν στο πεδίο EMAIL το κενό ' '. Ο κώδικας SQL που εκτελεί τα παραπάνω είναι:

```
insert
into db2admin.JS_WI_WCT_RNS_V2 (
select *
from db2admin.JS_WI_WCT_RNS
where email not like ' '
);
```

Τα νέα δεδομένα διαμορφώνονται ως εξής:

Πίνακας 5.6. Πλήθος εγγραφών με βάση το session ανά έτος (JS_WI_WCT_RNS_V2)

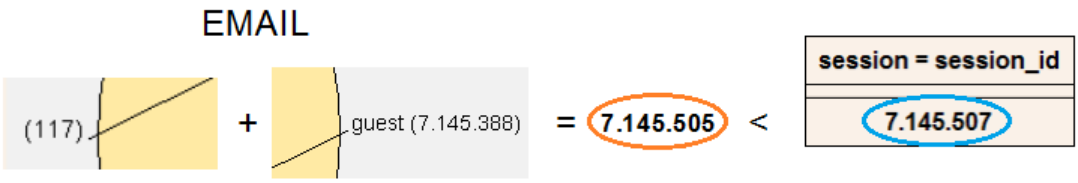
Έτος	session = null	session = 'ok'	session = session_id	Σύνολο
2005	0	4.823	118.212	123.035
2006	0	17.365	1.049.336	1.066.701
2007	0	15.873	2.515.786	2.531.659
2008	0	14.863	3.368.906	3.383.769
2009	0	2.878	93.150	96.028
Σύνολο	**0**	**55.802**	**7.145.390**	**7.201.192**

Πίνακας 5.7. Πλήθος μοναδικών συναλλαγών στον πίνακα JS_WI_WCT_RNS_V2

Συναλλαγές	Πλήθος
Μελών	17.386
Επισκεπτών	5.834.614

4. Δύο κρυμμένες εγγραφές

Εκτός από τα παραπάνω, το σχήμα 5.11 μας δίνει μία επιπλέον πληροφορία. Αθροίζοντας το πλήθος των εγγραφών που έχουν στο πεδίο EMAIL την τιμή 'guest' και το πλήθος των εγγραφών που έχουν την τιμή ' ', προκύπτει ο αριθμός 7.145.505. Συγκρίνοντας τον αριθμό αυτό με το πλήθος των εγγραφών του πίνακα JS_WI_WCT_RNS που στο πεδίο SESSION έχουν ένα μοναδικό session id (πίνακας 5.4), διαπιστώνουμε ότι διαφέρουν κατά δύο.



Σχήμα 5.12. Ασυμφωνία συνόλων εγγραφών που καταχωρήθηκαν από επισκέπτες

Από τη στιγμή που οι δύο αυτοί αριθμοί δε συμφωνούν, τότε σίγουρα υπάρχουν στον πίνακα JOURNAL_STATS εγγραφές με πραγματικό e-mail στο πεδίο EMAIL και με session id στο πεδίο SESSION. Εγγραφές δηλαδή που

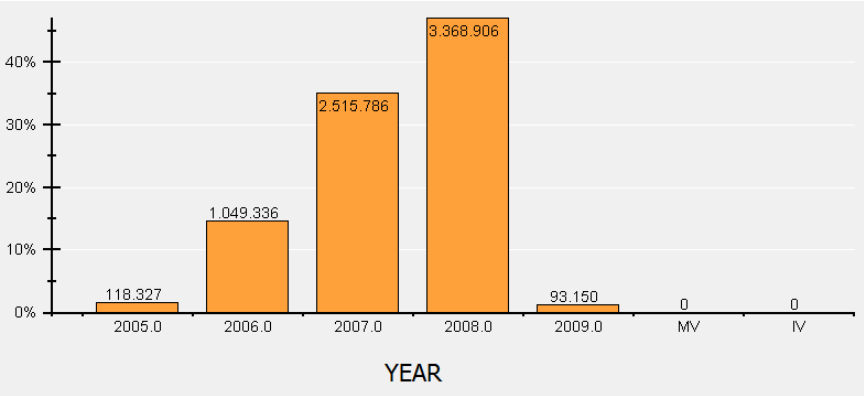
δε συμφωνούν με τους κανόνες σύμφωνα με τους οποίους λειτουργεί το σύστημα του HEAL-Link. Τις εγγραφές αυτές πρέπει να τις εντοπίσουμε και να τις αφαιρέσουμε.

Στην προσπάθειά μας να τις εντοπίσουμε, θα μας βοηθήσουν τρία στοιχεία.

Το πρώτο στοιχείο είναι ότι σίγουρα θα φέρουν στο πεδίο EMAIL πραγματικές διευθύνσεις. Είμαστε βέβαιοι για αυτό διότι από τη στιγμή που δεν συμμετέχουν στα αποτελέσματα που επέστρεψε η συνάρτηση “Univariate Distribution”, αυτό σημαίνει ότι εμποδίστηκαν από τον περιορισμό που απέκλειε από τη συνάρτηση τις εγγραφές εκείνες που έχουν πραγματικές διευθύνσεις στο πεδίο EMAIL ("JS_WI_WCT_RNS"."EMAIL" NOT LIKE '%@%').

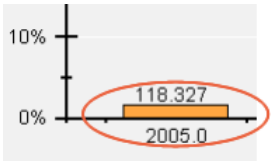
Το δεύτερο στοιχείο είναι η τιμή του πεδίου SESSION των αγνοούμενων εγγραφών. Το σύνολο των εγγραφών που έχουν session id στο πεδίο SESSION είναι κατά δύο (2) μεγαλύτερο από το σύνολο των εγγραφών που δεν έχουν πραγματικό e-mail στο πεδίο EMAIL. Το πλεόνασμα των εγγραφών ανήκει στην ομάδα εγγραφών όπου ισχύει session=session_id. Κατά συνέπεια, οι εγγραφές που αναζητάμε θα πρέπει να έχουν session id στο πεδίο SESSION.

Το τρίτο στοιχείο το αντλούμε από την παρατήρηση του πρώτου γραφήματος που επέστρεψε η συνάρτηση “Univariate Distribution”. Το γράφημα αυτό παρουσιάζει την κατανομή των τιμών στη στήλη YEAR. Δηλαδή, για κάθε έτος, πόσες εγγραφές από το σύνολο των εγγραφών που συμμετείχαν στην ανάλυση φέρουν στο πεδίο YEAR το συγκεκριμένο έτος.



Σχήμα 5.13. Πλήθος εγγραφών χωρίς πραγματικό EMAIL ανά έτος

Εάν συγκρίνουμε όλες αυτές τις τιμές με τις αντίστοιχες της στήλης session=session_id του πίνακα 5.4, θα διαπιστώσουμε ότι είναι όλες ίσες εκτός από τις τιμές του έτους 2005. Η συνάρτηση “Univariate Distribution” εντόπισε 118.327 εγγραφές καταχωρημένες το έτος 2005 ενώ ο πίνακας 5.4 μας ενημερώνει ότι το έτος 2005 καταχωρήθηκαν 118.329 εγγραφές. Και πάλι η διαφορά είναι ίση με δύο. Είναι φανερό λοιπόν ότι οι δύο εγγραφές τις οποίες αναζητάμε σίγουρα θα έχουν στο πεδίο YEAR την τιμή 2005.



Έτος	session = null	session = 'ok'	session = session_id	Σύνολο
2005	0	4.823	118.329	123.152

Σχήμα 5.14. Ασυμφωνία πλήθους εγγραφών που καταχωρήθηκαν από επισκέπτες για το έτος 2005

Με βάση τα τρία παραπάνω στοιχεία, μπορούμε να θέσουμε στη βάση δεδομένων το ακόλουθο SQL ερώτημα ώστε να μας δώσει τις αγνοούμενες εγγραφές:

```
select *
from db2admin.JS_WI_WCT_RNS_V2
where email like '%@%' and session not like 'ok' and year = 2005;
```

Η DB2 επέστρεψε τα ακόλουθα:

	ID	EMAIL	IP	TYPE	JOURNAL	SESSION	YEAR	MONTH	DAY	HOUR	MINUTE	SECOND	DISTANCE
1	208144	[redacted]@it.teithe.gr	155.207.114.8	alpha	E-Journal o...	8C10D200...	2005	10	17	13	59	20	1129546760000
2	208371	[redacted]@it.teithe.gr	127.0.0.1	alpha	Baby Talk	DA993399...	2005	10	17	14	54	31	1129550071000

Σχήμα 5.15. Οι δύο αγνοούμενες εγγραφές

Οι εγγραφές αυτές δε συμφωνούν με τους κανόνες λειτουργίας του συστήματος και μπορούν να χαρακτηριστούν ως έκτοπα. Συνεπώς, θα πρέπει να αφαιρεθούν από τα τελικά δεδομένα που θα τροφοδοτήσουμε στους αλγόριθμους εξόρυξης.

Δημιουργούμε ένα νέο πίνακα με όνομα JS_WI_WCT_RNS_V3 (version 3) και αντιγράφουμε σε αυτόν όλα τα περιεχόμενα του πίνακα JS_WI_WCT_RNS_V2 εξαιρώντας τις δύο εγγραφές που θέλουμε να αφαιρέσουμε. Ο κώδικας SQL που εκτελεί η DB2 είναι:

```
insert
into db2admin.JS_WI_WCT_RNS_V3 (
select *
from db2admin.JS_WI_WCT_RNS_V2
where id <> 208144 and id <> 208371
);
```

Τα νέα δεδομένα διαμορφώνονται ως εξής:

Πίνακας 5.8. Πλήθος εγγραφών με βάση το session ανά έτος (JS_WI_WCT_RNS_V3)

Έτος	session = null	session = 'ok'	session = session_id	Σύνολο
2005	0	4.823	118.210	123.033
2006	0	17.365	1.049.336	1.066.701
2007	0	15.873	2.515.786	2.531.659
2008	0	14.863	3.368.906	3.383.769
2009	0	2.878	93.150	96.028
Σύνολο	**0**	**55.802**	**7.145.388**	**7.201.190**

Πίνακας 5.9. Πλήθος μοναδικών συναλλαγών στον πίνακα JS_WI_WCT_RNS_V3

Συναλλαγές	Πλήθος
Μελών	17.386
Επισκεπτών	5.834.612

5. Συναλλαγές που εκτελέστηκαν από Web Crawlers

Οι Web Crawlers είναι προγράμματα υπολογιστών τα οποία περιηγούνται στον παγκόσμιο ιστό με μεθοδευμένο και αυτοματοποιημένο τρόπο με σκοπό την άντληση πληροφοριών. Πολλοί διαδικτυακοί τόποι, κυρίως μηχανές αναζήτησης, χρησιμοποιούν τους web crawlers για να δημιουργήσουν αντίγραφα των ιστοσελίδων που επισκέπτονται ώστε να τις ευρετηριοποιήσουν και να είναι ταχύτερη η μετέπειτα αναζήτησή τους. Επίσης, μπορούν να χρησιμοποιηθούν για λειτουργίες συντήρησης ιστοσελίδων όπως αναζήτηση σπασμένων υπερσυνδέσμων και επικύρωση του κώδικα HTML. Μία λιγότερο καλόβουλη χρήση τους, είναι αυτή της αναζήτησης συγκεκριμένων πληροφοριών από το διαδίκτυο όπως διευθύνσεις ηλεκτρονικού ταχυδρομείου οι οποίες πολλές φορές χρησιμοποιούνται για spamming.

Οι crawlers δεν άφησαν ανεπηρέαστο ούτε τον διαδικτυακό τόπο του HEAL-Link. Για την ακρίβεια, τον επισκέφτηκαν και μάλιστα πολλές χιλιάδες φορές σύμφωνα με τις πληροφορίες που είναι καταγεγραμμένες στη βάση δεδομένων.

Όταν ένας crawler εισέρχεται στον διαδικτυακό τόπο του HEAL-Link, τότε ξεκινάει μία νέα συνεδρία με την οποία αντιστοιχίζεται (το σύστημα τον αντιλαμβάνεται σαν κανονικό χρήστη). Ο crawler διεισδύει με πολύ γρήγορους ρυθμούς σε όλους τους υπερσυνδέσμους του διαδικτυακού τόπου. Κατά συνέπεια, διεισδύει και στους υπερσυνδέσμους του καταλόγου και των θεματικών ενοτήτων φτάνοντας έτσι στα συγγράμματα. Κάθε φορά που ακολουθεί τον υπερσύνδεσμο ενός συγγράμματος, το σύστημα προσθέτει μία νέα εγγραφή στον πίνακα JOURNAL_STATS. Κατά αυτόν τον τρόπο, καταχωρούνται σε σύντομο χρονικό διάστημα χιλιάδες λήψεις συγγραμμάτων εντός μίας μόνο συναλλαγής, κάτι το οποίο είναι απίθανο να συμβεί από άνθρωπο.

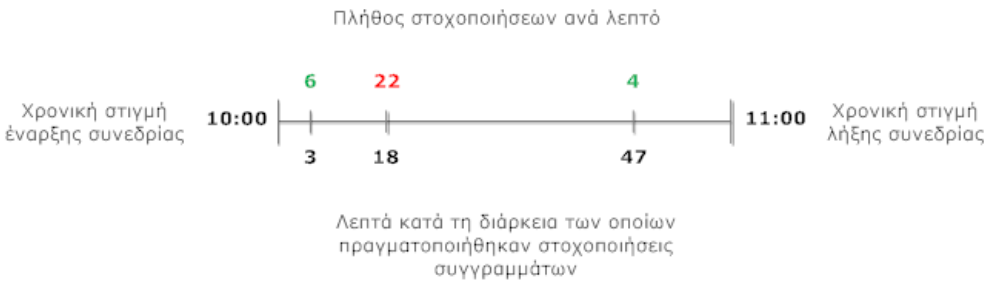
Σκοπός μας είναι να εντοπίσουμε τις συναλλαγές που εκτελέστηκαν από web crawlers, να τις μελετήσουμε και να τις αφαιρέσουμε από τα τελικά δεδομένα. Η προσπάθειά μας θα εκτελεστεί σε τρία βήματα.

5.1. Ποιες συναλλαγές εκτελέστηκαν από web crawlers;

Οι συναλλαγές οι οποίες εκτελέστηκαν από web crawlers περιλαμβάνουν πολύ μεγάλο αριθμό επισκέψεων σε συγγράμματα σε πολύ μικρά χρονικά διαστήματα. Για να τις εντοπίσουμε, θα μετρήσουμε τις επισκέψεις που έγιναν σε κάθε συναλλαγή ανά λεπτό. Η καταμέτρηση δε θα γίνει για όλα τα λεπτά της ώρας (1-60), αλλά μόνο για τα λεπτά κατά τη διάρκεια των οποίων έγιναν επισκέψεις. Για παράδειγμα, εάν σε μία συναλλαγή καταγράφηκαν επισκέψεις στα λεπτά 3, 18 και 47, τότε θα μετρήσουμε το πλήθος των επισκέψεων που έγιναν μόνο σε αυτά τα λεπτά. Επειδή είναι πιθανό να καταγράφηκαν επισκέψεις στο ίδιο λεπτό αλλά σε διαφορετικές ώρες (π.χ. 14:03 και 15:03), θα πρέπει να λάβουμε υπόψιν και την ώρα, ώστε να αποφύγουμε την άθροιση του πλήθους των επισκέψεων που έγιναν στο ίδιο λεπτό σε διαφορετικές ώρες (θέλουμε το πλήθος των επισκέψεων ανά ξεχωριστό λεπτό).

Στη συνέχεια, θα πρέπει να θέσουμε ένα όριο επισκέψεων ανά λεπτό πέραν του οποίου θα θεωρούμε ότι η συναλλαγή εκτελέστηκε από web crawler και όχι από άνθρωπο. Ορίζουμε το όριο αυτό στην τιμή 15. Δηλαδή, όσες συναλλαγές έχουν περισσότερες από 15 επισκέψεις σε συγγράμματα σε κάποιο λεπτό της διάρκειάς τους, θεωρούμε ότι εκτελέστηκαν από web crawlers.

Στο ακόλουθο διάγραμμα παρουσιάζεται η παραπάνω ιδέα:



Σχήμα 5.16. Πλήθος στοχοποιήσεων ανά λεπτό κατά τη διάρκεια μίας συνεδρίας

Παρατηρούμε στο διάγραμμα ότι στη συνεδρία που ξεκίνησε στις 10:00 και ολοκληρώθηκε στις 11:00 πραγματοποιήθηκαν στοχοποιήσεις συγγραμμάτων στα λεπτά 3, 18 και 47. Μετρώντας το πλήθος των στοχοποιήσεων που έγιναν σε κάθε λεπτό διαπιστώνουμε ότι, στο 3ο λεπτό έγιναν 6, στο 18ο 22 και στο 47ο 4. Επειδή υπάρχει τουλάχιστον ένα λεπτό κατά τη διάρκεια του οποίου έγιναν περισσότερες από 15 στοχοποιήσεις (ξεπερνώντας το όριο που θέσαμε πιο πάνω), θεωρούμε ότι η συναλλαγή αυτή εκτελέστηκε από web crawler.

Να σημειώσουμε σε αυτό το σημείο ότι η λύση που ακολουθούμε ενδεχομένως να μην είναι η ιδανική. Μπορεί με τη συγκεκριμένη προσέγγιση του προβλήματος να αφαιρούμε πολλά crawler-sessions, όχι όμως όλα. Δεδομένου ότι η διεργασία εξόρυξης που εφαρμόζουμε είναι επαναληπτική, μπορούμε να επιστρέψουμε στο τρέχον πρόβλημα σε μεταγενέστερο στάδιο και να εκτελέσουμε περαιτέρω ενέργειες καθαρισμού των δεδομένων εάν διαπιστώσουμε ότι τα παραγόμενα μοντέλα κανόνων συσχετίσεων δε μας ικανοποιούν.

Για να μάθουμε ποιες συναλλαγές εκτελέστηκαν από web crawlers εκτελούμε τον ακόλουθο κώδικα:

```
select distinct session
from db2admin.JS_WI_WCT_RNS_V3
where session not like 'ok'
```

```
group by session, hour, minute
having count(*) > 15
```

Η DB2 θα ομαδοποιήσει τις εγγραφές του πίνακα JS_WI_WCT_RNS_V3 με βάση τα πεδία SESSION, HOUR και MINUTE (εξαιρώντας τις εγγραφές που έχουν στο πεδίο SESSION την τιμή 'ok') και θα κρατήσει εκείνες τις ομάδες (groups) που έχουν πλήθος μεγαλύτερο από 15. Εγγραφές που καταχωρήθηκαν στο ίδιο λεπτό, της ίδιας ώρας, στην ίδια συναλλαγή, θα τοποθετηθούν στην ίδια ομάδα. Μπορεί να υπάρχουν συναλλαγές στις οποίες να καταγράφηκαν περισσότερες από 15 επισκέψεις συγγραμμάτων σε περισσότερα από ένα λεπτά. Σε αυτές τις περιπτώσεις, οι εν λόγω συναλλαγές θα συμμετέχουν στο τελικό αποτέλεσμα περισσότερες από μία φορά. Για να είμαστε σίγουροι ότι θα εμφανιστούν μόνο μία φορά, χρησιμοποιούμε την εντολή “distinct”.

Για να μάθουμε το πλήθος των συναλλαγών αυτών, εκτελούμε:

```
select count(*)
from (
select distinct session
from db2admin.JS_WI_WCT_RNS_V3
where session not like 'ok'
group by session, hour, minute
having count(*) > 15
) groups;
```

Το αποτέλεσμα που επιστρέφεται είναι 85. Δηλαδή, στον πίνακα JS_WI_WCT_RNS_V3 είναι καταχωρημένες 85 μοναδικές συναλλαγές στις οποίες έγιναν περισσότερες από 15 επισκέψεις σε συγγράμματα σε κάποιο ή κάποια από τα λεπτά της διάρκειάς τους.

Πόσες όμως επισκέψεις έγιναν σε κάθε ένα από αυτά τα λεπτά ξεχωριστά; Η απάντηση σε αυτό το ερώτημα θα μας δώσει μία εικόνα της επίδρασης που είχαν οι web crawlers στον δικτυακό τόπο του HEAL-Link.

5.2. Πόσες επισκέψεις έγιναν σε κάθε ένα από τα λεπτά των μοναδικών συναλλαγών;

Για να απαντηθεί αυτό το ερώτημα δεν έχουμε παρά να εκτελέσουμε τον ακόλουθο SQL κώδικα:

```
select session, hour, minute, count(*) as visits
from db2admin.JS_WI_WCT_RNS_V3
where session not like 'ok'
group by session, hour, minute
having count(*) > 15
order by visits desc;
```

Το SQL ερώτημα είναι ίδιο με αυτά που εκτελέστηκαν στο προηγούμενο βήμα, με μόνη διαφορά το επιστρεφόμενο αποτέλεσμα. Επιστρέφονται οι ομάδες ταξινομημένες κατά φθίνουσα σειρά με βάση τη στήλη VISITS. Υπενθυμίζουμε ότι, κάθε ομάδα εκπροσωπεί ένα συγκεκριμένο λεπτό, μίας συγκεκριμένης ώρας, μίας συγκεκριμένης συναλλαγής. Στο ακόλουθο σχήμα παρουσιάζονται οι 40 πρώτες ομάδες. Ας προσπαθήσουμε να διαβάσουμε την πρώτη και την δεύτερη γραμμή:

1η γραμμή: Στο δωδέκατο λεπτό, της ενδέκατης ώρας, της συνεδρίας (συναλλαγής) με session = session_id, έγιναν 225 επισκέψεις.

2η γραμμή: Στο πεντηκοστό δεύτερο λεπτό, της εικοστής τέταρτης ώρας, της συνεδρίας (συναλλαγής) με session = session_id, έγιναν 219 επισκέψεις.

Παρόλο που στην εικόνα φαίνονται τα νούμερα 10, 11, 23 και 51, εμείς διαβάσαμε αντίστοιχα 11, 12, 24 και 52 αφού η μέτρηση των ωρών και των λεπτών ξεκινάει από το 0. Για παράδειγμα, εάν μία επίσκεψη έγινε στο πρώτο λεπτό της πρώτης ώρας του εικοσιτετραώρου, τότε στα πεδία HOUR και MINUTE θα ήταν καταχωρημένη η τιμή μηδέν.

	SESSION	HOUR	MINUTE	VISITS
1	9AA6C9F24ED5102DE2CDA601CBE6211C	10	11	225
2	6443A257688486D83DDCABAACE0BA1E9	23	51	219
3	9AA6C9F24ED5102DE2CDA601CBE6211C	10	14	219
4	8D04517AE9196EEA2036615E7C54223F	8	51	218
5	E908E82E4F2D14757A59FCEC522B7293	7	40	218
6	9AA6C9F24ED5102DE2CDA601CBE6211C	10	13	217
7	0F49C65B86CBE07BFE5C2703C2466B6E	16	36	215
8	E64D617B6AA9F8493EE563D0B25E5E1F	16	9	215
9	901C73B640105BF4067551969B5B4DE3	14	29	212
10	0F49C65B86CBE07BFE5C2703C2466B6E	16	40	209
11	8D04517AE9196EEA2036615E7C54223F	8	36	209
12	429F3CE0FC66BD1BC5632861A3A747A0	10	17	208
13	8D04517AE9196EEA2036615E7C54223F	8	32	208
14	8D04517AE9196EEA2036615E7C54223F	8	47	208
15	8D04517AE9196EEA2036615E7C54223F	8	50	208
16	8D04517AE9196EEA2036615E7C54223F	8	52	208
17	8D04517AE9196EEA2036615E7C54223F	8	34	206
18	8D04517AE9196EEA2036615E7C54223F	8	35	206
19	8D04517AE9196EEA2036615E7C54223F	8	39	206
20	4BEFE7FFB05204DF2BFD460B71182788	19	59	205
21	8D04517AE9196EEA2036615E7C54223F	8	38	204
22	8D04517AE9196EEA2036615E7C54223F	8	33	203
23	8D04517AE9196EEA2036615E7C54223F	8	42	203
24	9AA6C9F24ED5102DE2CDA601CBE6211C	10	12	203
25	9AA6C9F24ED5102DE2CDA601CBE6211C	10	15	202
26	E64D617B6AA9F8493EE563D0B25E5E1F	16	8	201
27	8D04517AE9196EEA2036615E7C54223F	8	37	199
28	A08FEC0ACC36A506F66738946FF4CFE1	20	4	199
29	8D04517AE9196EEA2036615E7C54223F	8	40	198
30	429F3CE0FC66BD1BC5632861A3A747A0	10	15	196
31	8D04517AE9196EEA2036615E7C54223F	8	43	193
32	684622FF9C9685344D7D68B874334F33	12	11	192
33	8D04517AE9196EEA2036615E7C54223F	8	30	192
34	8D04517AE9196EEA2036615E7C54223F	8	44	192
35	901C73B640105BF4067551969B5B4DE3	14	30	191
36	429F3CE0FC66BD1BC5632861A3A747A0	10	20	190
37	429F3CE0FC66BD1BC5632861A3A747A0	10	16	187
38	6443A257688486D83DDCABAACE0BA1E9	23	53	185
39	8D04517AE9196EEA2036615E7C54223F	8	48	185
40	8D04517AE9196EEA2036615E7C54223F	8	31	184

Σχήμα 5.17. Μερικά από τα λεπτά κατά τη διάρκεια των οποίων έγιναν περισσότερες από 15 επισκέψεις. Ταξινομημένα κατά φθίνουσα σειρά με βάση το πλήθος των επισκέψεων.

Ένα δείγμα των ίδιων αποτελεσμάτων ομαδοποιημένα με βάση τη στήλη SESSION φαίνεται στο ακόλουθο σχήμα. Είναι εύκολο να παρατηρήσουμε ότι, εντός της ίδιας συναλλαγής (τέταρτη κατά σειρά), υπάρχουν λεπτά τα οποία επαναλαμβάνονται σε διαφορετικές ώρες (π.χ. 1ο λεπτό, 2ο λεπτό, 3ο λεπτό, κλπ). Εάν στην ομαδοποίηση δεν χρησιμοποιούσαμε το πεδίο HOUR (group by session, hour, minute), τότε η DB2 θα εμφάνιζε κάθε επαναλαμβανόμενο λεπτό μία μόνο φορά αθροίζοντας τα σύνολα επισκέψεων κάθε επανάληψης. Για παράδειγμα, οι γραμμές 4 και 5 θα γινόταν μία γραμμή με πλήθος επισκέψεων για το πρώτο λεπτό (minute 0) 99+70.

	SESSION	HOUR	MINUTE	VISITS
1	006E1B268EE6E49426F974F7EDBEC876	8	34	59
2	01E90403F42DC1B11941973CB655D359	1	30	17
3	01E90403F42DC1B11941973CB655D359	1	31	18
4	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	0	99
5	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	0	70
6	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	1	110
7	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	1	71
8	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	2	110
9	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	2	69
10	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	3	110
11	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	3	71
12	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	4	104
13	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	4	70
14	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	5	111
15	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	5	71
16	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	6	119
17	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	6	71
18	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	7	110
19	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	7	72
20	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	8	109
21	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	8	76
22	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	9	104
23	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	9	71
24	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	10	103
25	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	10	70
26	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	11	97
27	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	11	72
28	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	12	105
29	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	12	74
30	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	13	104
31	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	13	78
32	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	14	113
33	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	14	72
34	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	15	101
35	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	15	87
36	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	16	106
37	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	16	83
38	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	17	105
39	04EB204A3AAB35CA9EFD9FA7249A4E7C	5	17	79
40	04EB204A3AAB35CA9EFD9FA7249A4E7C	4	18	84

Σχήμα 5.18. Μερικά από τα λεπτά κατά τη διάρκεια των οποίων έγιναν περισσότερες από 15 επισκέψεις. Ομαδοποιημένα κατά SESSION.

Στο ακόλουθο σχήμα παρουσιάζονται με περισσότερες λεπτομέρειες 30 από τις 85 μοναδικές συναλλαγές που εκτελέστηκαν από crawlers καθώς επίσης και ο κώδικας SQL που παρήγαγε τα αποτελέσματα. Για κάθε συναλλαγή φαίνεται η IP προέλευσης του crawler, οι χρονικές στιγμές έναρξης και λήξης της συναλλαγής και το πλήθος των επισκέψεων που εκτέλεσε ο crawler κατά τη διάρκειά της.

```
select session, max(ip) as IP, timestamp('1970-01-01','00.00.00') + (min(distance)/1000) seconds,
                               timestamp('1970-01-01','00.00.00') + (max(distance)/1000) seconds, count(*) as visits
from (
   select session, ip, distance
   from db2admin.JS_WI_WCT_RNS_V3
   where session in (
      select distinct session
      from db2admin.JS_WI_WCT_RNS_V3
      where session not like 'ok'
      group by session, hour, minute
      having count(*) > 15
   )
) as tmp
group by session;
```

A/A	SESSION	IP	STARTED_AT	FINISHED_AT	VISITS
1	006E1B268EE6E49426F974F7EDBEC876	89.122.29.122	2008-05-12 05:33:29.0	2008-05-12 05:38:47.0	70
2	01E90403F42DC1B11941973CB655D359	89.137.233.185	2007-02-26 23:29:03.0	2007-02-26 23:32:47.0	61
3	04EB204A3AAB35CA9EFD9FA7249A4E7C	212.205.108.122	2006-03-11 01:39:04.0	2006-03-11 03:37:27.0	10343
4	056A080D0EDB9960E20AA10826865F01	212.205.108.122	2006-03-10 01:40:17.0	2006-03-10 03:53:54.0	10361
5	0620126E6CA3EEF40B55BEC4492BF9B3	62.103.55.146	2006-05-06 15:45:12.0	2006-05-07 09:50:09.0	10037
6	07FE2F3872A905E0BC430450774EF7E9	62.103.55.146	2006-05-13 22:00:29.0	2006-05-14 21:46:42.0	10361
7	090893D16AB9AD16F3788F626B3FD7F3	212.205.108.122	2006-03-01 16:21:38.0	2006-03-02 10:56:39.0	1529
8	09A6FD3E6E6729021A4BDA994FD7CC42	62.103.213.114	2009-02-27 09:47:44.0	2009-02-27 09:47:52.0	16
9	0F49C65B86CBE07BFE5C2703C2466B6E	139.91.254.18	2007-06-27 13:34:49.0	2007-06-27 13:42:28.0	1354
10	167A32A6A909F42B8B1F599BB9E0051C	212.205.108.122	2006-03-23 01:39:37.0	2006-03-23 03:27:43.0	10335
11	185B6B8C2639615BBCE9CCE417187D7A	87.202.126.171	2008-01-08 06:07:37.0	2008-01-08 06:08:04.0	70
12	1B77281E65B3B0015B04445E3ED0D286	212.205.108.122	2006-03-01 11:06:11.0	2006-03-01 22:58:43.0	362
13	1F3DB3D4141ABFA86B1E07AD91DD8C36	72.215.220.52	2007-10-25 03:09:27.0	2007-10-25 03:14:30.0	273
14	20A2C0AFF36FBF0419E112106180F436	62.103.55.146	2006-04-12 20:36:04.0	2006-04-13 09:59:59.0	10358
15	23B9BF2FCCEA8EBA2081DF9B29EB805B	62.103.55.146	2006-02-17 23:40:11.0	2006-02-18 22:59:57.0	8876
16	27B46878E52075DEBDC614E4C5192FE1	62.103.55.146	2006-04-13 23:37:24.0	2006-04-14 01:46:26.0	10338
17	2868E2506CE0E2F78B3B08775384FB2D	212.205.108.122	2006-03-28 06:34:41.0	2006-03-28 07:13:24.0	2604
18	2AD64D3A1E90749AE68DB93DABDFA896	147.52.249.146	2005-11-11 14:31:37.0	2005-11-11 14:45:51.0	21
19	2F7ADF8C218C714C3125D0AB5FD13D8C	85.75.15.157	2009-02-28 14:28:13.0	2009-02-28 14:28:26.0	35
20	3EC10F69DEB14F3FA17AA3BF7C103D2B	149.254.200.215	2008-02-24 22:31:15.0	2008-02-24 22:33:29.0	40
21	41F5F77AB45451C5F50D06E13F80848C	91.140.21.145	2007-04-12 22:10:52.0	2007-04-13 09:46:40.0	62
22	429F3CE0FC66BD1BC5632861A3A747A0	212.205.108.122	2006-03-27 06:07:02.0	2006-03-27 07:37:09.0	7355
23	4326B76AA7650554B03F3E360D5D2729	212.205.108.122	2006-03-27 23:40:35.0	2006-03-28 00:39:17.0	4260
24	4BEFE7FFB05204DF2BFD460B71182788	222.46.88.18	2005-12-17 17:54:29.0	2005-12-17 18:00:15.0	590
25	4DA56DF0B0EEE920D281A83E7F410164	62.85.45.65	2007-12-15 14:25:51.0	2007-12-15 14:27:42.0	30
26	5A4B0CDADF2CC2C6B753F2E4C23A85AC	147.83.153.155	2006-09-10 02:15:10.0	2006-09-10 02:23:01.0	168
27	5AD581C7A12B0959130C58ABBDD1CE5A	79.166.112.79	2009-01-07 00:05:08.0	2009-01-07 00:05:44.0	35
28	5D0DBA8C934325CE56635517AA35EBD0	147.83.153.156	2006-10-07 05:55:36.0	2006-10-07 06:08:52.0	168
29	5DB3F18FD81F6BDE0D62825E8166603D	67.162.215.173	2008-01-31 18:16:48.0	2008-01-31 18:26:18.0	41
30	5F80D67C93DA8B35AE10C91AF27ED051	62.103.26.228	2007-04-16 15:11:24.0	2007-04-16 15:12:17.0	71

Σχήμα 5.19. Δείγμα των συναλλαγών που εκτελέστηκαν από Web Crawlers.

Όπως φαίνεται από τους παραπάνω πίνακες, οι crawlers που εισήλθαν στον διαδικτυακό τόπο του HEAL-Link ήταν αρκετά δραστήριοι. Ορισμένοι από αυτούς ξεπέρασαν τις 200 επισκέψεις συγγραμμάτων ανά λεπτό. Οι χρόνοι, βέβαια, στους οποίους ολοκλήρωσαν το έργο τους ποικίλουν, γεγονός το οποίο πιθανόν οφείλεται στην ποιότητα του κώδικα κάθε crawler.

Αθροίζοντας το πλήθος των επισκέψεων σε όλες τις συναλλαγές που έγιναν από crawlers προκύπτει ο αριθμός 241.125. Αυτό είναι το συνολικό πλήθος των εγγραφών που θα αφαιρεθούν από τον πίνακα JS_WI_WCT_RNS_V3 στο αμέσως επόμενο βήμα.

5.3. Αφαίρεση των συναλλαγών που εκτελέστηκαν από web crawlers

Σε αυτό το βήμα, θα αφαιρέσουμε από τα δεδομένα όλες τις συναλλαγές που εκτελέστηκαν από web crawlers. Δημιουργούμε ένα νέο πίνακα με όνομα JS_WI_WCT_RNS_RCT (Journal-Stats-With-ID-With-Converted-Timestamps-Removed-Null-Sessions-Removed-Crawler-Transactions) και αντιγράφουμε σε αυτόν όλα τα δεδομένα του πίνακα JS_WI_WCT_RNS_V3, εξαιρώντας τις συναλλαγές που εκτελέστηκαν από crawlers.

Ο κώδικας αντιγραφής των δεδομένων είναι:

```
insert
into db2admin.JS_WI_WCT_RNS_RCT
select *
```

```
from db2admin.JS_WI_WCT_RNS_V3
where session not in (
select distinct session
from db2admin.JS_WI_WCT_RNS_V3
where session not like 'ok'
group by session, hour, minute
having count(*) > 15
);
```

Μετά την εκτέλεση του παραπάνω κώδικα, ο πίνακας JS_WI_WCT_RNS_RCT θα περιέχει όλα τα δεδομένα του JS_WI_WCT_RNS_V3 εκτός από τις 241.125 εγγραφές που καταχωρήθηκαν από web crawlers.

Τα νέα δεδομένα διαμορφώνονται ως εξής:

Πίνακας 5.10. Πλήθος εγγραφών με βάση το session ανά έτος (JS_WI_WCT_RNS_RCT)

Έτος	session = null	session = 'ok'	session = session_id	Σύνολο
2005	0	4.823	116.984	121.807
2006	0	17.365	838.025	855.390
2007	0	15.873	2.506.942	2.522.815
2008	0	14.863	3.349.317	3.364.180
2009	0	2.878	92.995	95.873
Σύνολο	**0**	**55.802**	**6.904.263**	**6.960.065**

Πίνακας 5.11. Πλήθος μοναδικών συναλλαγών στον πίνακα JS_WI_WCT_RNS_RCT

Συναλλαγές	Πλήθος
Μελών	17.386
Επισκεπτών	5.834.527

6. Στοχοποιήσεις συγγραμμάτων από τον localhost

Αναζητώντας στον πίνακα JS_WI_WCT_RNS_RCT εγγραφές που προήλθαν από τον τοπικό διακομιστή (δηλαδή από τον server που φιλοξενεί το σύστημα του HEAL-Link) εκτελούμε τον παρακάτω κώδικα:

```
select *
from db2admin.JS_WI_WCT_RNS_RCT
where IP like '127%';
```

και η DB2 επιστρέφει τα ακόλουθα αποτελέσματα:

	ID	EMAIL	IP	TYPE	JOURNAL	SESSION	YEAR	MONTH	DAY	HOUR	MINUTE	SECOND	DISTANCE
1	208344	[redacted]@it.teithe.gr	127.0.0.1	alpha	Dados	ok	2005	10	17	14	46	14	1129549...
2	208364	[redacted]@it.teithe.gr	127.0.0.1	alpha	Dados	ok	2005	10	17	14	53	20	1129550...
3	208365	[redacted]@it.teithe.gr	127.0.0.1	alpha	DELTA	ok	2005	10	17	14	53	26	1129550...
4	208351	[redacted]@it.teithe.gr	127.0.0.1	alpha	Facilities	ok	2005	10	17	14	49	18	1129549...
5	208382	[redacted]@it.teithe.gr	127.0.0.1	alpha	CA Maga...	ok	2005	10	17	14	58	50	1129550...
6	314048	[redacted]@physics.auth.gr	127.0.0.1	alpha	Comput...	ok	2005	12	12	10	38	31	1134376...
7	313985	[redacted]@physics.auth.gr	127.0.0.1	alpha	Informati...	ok	2005	12	12	10	28	22	1134376...
8	208314	guest	127.0.0.1	alpha	Datamati...	65BC7BD51...	2005	10	17	14	38	17	1129549...
9	208337	guest	127.0.0.1	alpha	Dados	88998929A...	2005	10	17	14	44	46	1129549...
10	208394	[redacted]@it.teithe.gr	127.0.0.1	alpha	General a...	ok	2005	10	17	14	59	55	1129550...
11	313959	[redacted]@physics.auth.gr	127.0.0.1	alpha	Learning ...	ok	2005	12	12	10	23	2	1134375...
12	313942	guest	127.0.0.1	alpha	3C ON-LI...	D0DC2D4F...	2005	12	12	10	18	28	1134375...
13	313983	[redacted]@physics.auth.gr	127.0.0.1	alpha	Comput...	ok	2005	12	12	10	28	14	1134376...
14	314042	[redacted]@physics.auth.gr	127.0.0.1	alpha	ACM SIG...	ok	2005	12	12	10	37	47	1134376...
15	314050	guest	127.0.0.1	alpha	Lecture ...	7F87F90EB7...	2005	12	12	10	38	55	1134376...

Σχήμα 5.20. Επισκέψεις σε συγγράμματα από τον localhost

Παρατηρούμε ότι υπάρχουν 15 εγγραφές που καταχωρήθηκαν από τον localhost (127.0.0.1). Οι εγγραφές αυτές προέκυψαν από δοκιμές που εκτέλεσαν οι διαχειριστές του συστήματος, οπότε δε φέρουν χρήσιμες πληροφορίες για τη συμπεριφορά των χρηστών. Συνεπώς, μπορούμε να τις αφαιρέσουμε.

Δημιουργούμε ένα νέο πίνακα με όνομα JS_WI_WCT_RNS_RCT_V2 και αντιγράφουμε σε αυτόν όλα τα δεδομένα του πίνακα JS_WI_WCT_RNS_RCT, εξαιρώντας τις 15 εγγραφές που προήλθαν από τον localhost.

```
insert
into db2admin.JS_WI_WCT_RNS_RCT_V2
(select *
from db2admin.JS_WI_WCT_RNS_RCT
where ip not like '127%'
);
```

Τα νέα δεδομένα διαμορφώνονται ως εξής:

Πίνακας 5.12. Πλήθος εγγραφών με βάση το session ανά έτος (JS_WI_WCT_RNS_RCT_V2)

Έτος	session = null	session = 'ok'	session = session_id	Σύνολο
2005	0	4.812	116.980	121.792
2006	0	17.365	838.025	855.390
2007	0	15.873	2.506.942	2.522.815
2008	0	14.863	3.349.317	3.364.180
2009	0	2.878	92.995	95.873
Σύνολο	**0**	**55.791**	**6.904.259**	**6.960.050**

Πίνακας 5.13. Πλήθος μοναδικών συναλλαγών στον πίνακα JS_WI_WCT_RNS_RCT_V2

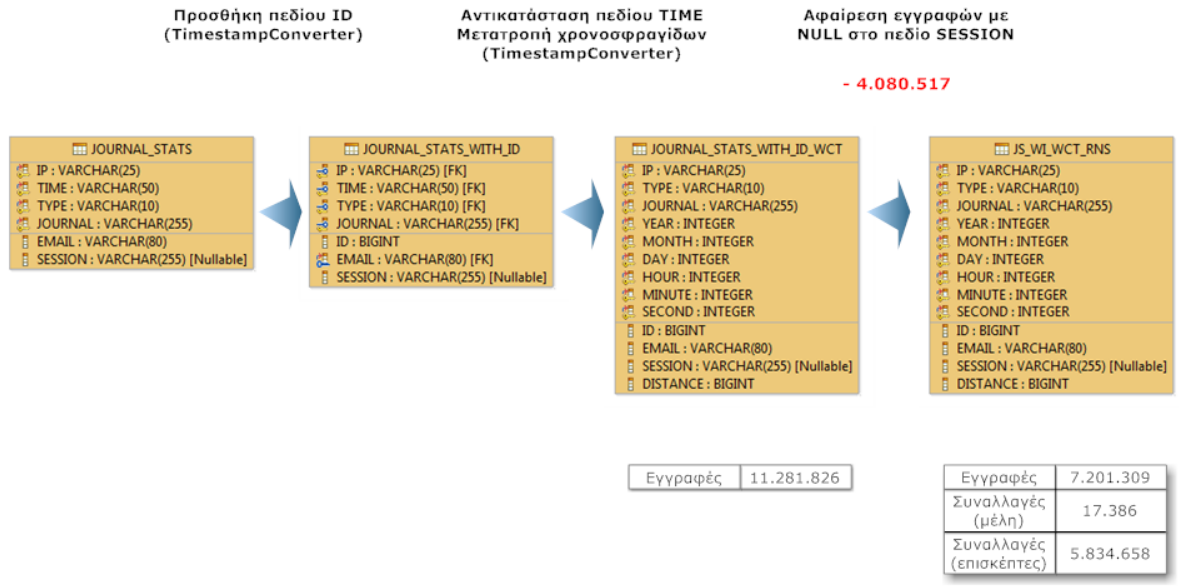
Συναλλαγές	Πλήθος
Μελών	17.386
Επισκεπτών	5.834.523

7. Ολοκλήρωση της προετοιμασίας των δεδομένων

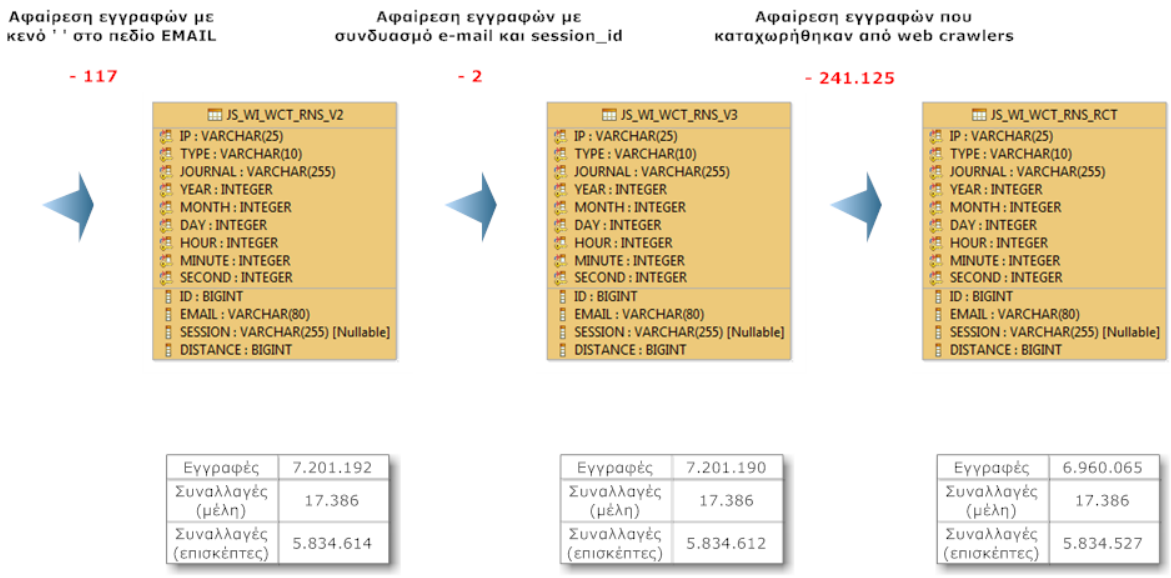
Με την αφαίρεση των συναλλαγών που εκτελέστηκαν από web crawlers, ολοκληρώνεται η πρώτη επανάληψη προετοιμασίας των δεδομένων. Ενδεχομένως να υπάρξουν και άλλες επαναλήψεις, ανάλογα με τις ανάγκες που θα προκύψουν στη φάση μοντελοποίησης που ακολουθεί.

Στην τελική μορφή στην οποία περιήλθε ο πίνακας συναλλαγών JS_WI_WCT_RNS_RCT_V2 περιλαμβάνει συνολικά 6.960.050 εγγραφές, από 11.281.826 που περιείχε αρχικά ο πίνακας JOURNAL_STATS. Το πλήθος των μοναδικών συναλλαγών για τα εγγεγραμμένα μέλη ανέρχεται σε 17.386 ενώ για τους επισκέπτες σε 5.834.523.

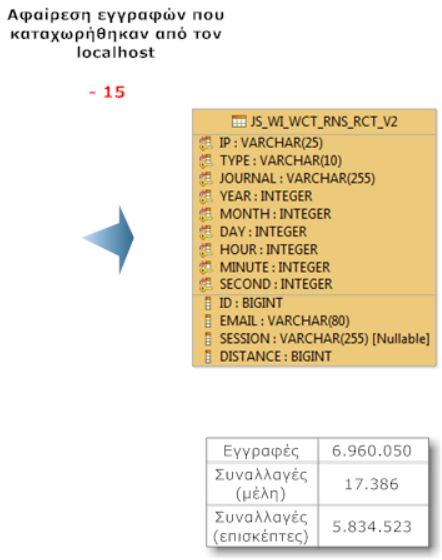
Στα ακόλουθα σχήματα συνοψίζονται όλες οι ενέργειες προετοιμασίας των δεδομένων που εκτελέστηκαν μέχρι αυτό το σημείο.



Σχήμα 5.21. Ιστορικό αλλαγών του πίνακα JOURNAL_STATS (1/3)



Σχήμα 5.22. Ιστορικό αλλαγών του πίνακα JOURNAL_STATS (2/3)



Σχήμα 5.23. Ιστορικό αλλαγών του πίνακα JOURNAL_STATS (3/3)

Τα δεδομένα βρίσκονται πλέον σε αρκετά ικανοποιητικό επίπεδο σε ότι αφορά την προετοιμασία τους πριν τροφοδοτηθούν στον αλγόριθμο εξόρυξης. Στο επόμενο κεφάλαιο ξεκινάμε τη διαδικασία μοντελοποίησης. Θα χρησιμοποιήσουμε το Design Studio για να δημιουργήσουμε και να εκτελέσουμε τα μοντέλα εξόρυξης, αφού πρώτα φέρουμε τα δεδομένα σε διάταξη συναλλαγών (transactional layout), και τον IM Visualizer για να μελετήσουμε και να αξιολογήσουμε τους κανόνες συσχετίσεων.

Κεφάλαιο 6. Μοντελοποίηση και αξιολόγηση

Στο κεφάλαιο αυτό θα δημιουργήσουμε τα μοντέλα συσχετίσεων τα οποία θα μας βοηθήσουν να ανακαλύψουμε κρυμμένες συσχετίσεις μεταξύ των αντικειμένων που βρίσκονται αποθηκευμένα στη βάση δεδομένων (ψηφιακά συγγράμματα, θεματικοί όροι, υποκατηγορίες και κατηγορίες θεματικών όρων).

Η δημιουργία των μοντέλων παραγωγής κανόνων συσχετίσεων θα γίνει εντός του Design Studio ενώ για την προβολή και αξιολόγησή τους θα χρησιμοποιήσουμε τον Intelligent Miner Visualizer.

1. Δύο διαφορετικά μοντέλα εξόρυξης (επισκεπτών και μελών)

Λόγω της φύσης των δεδομένων του διαδικτυακού τόπου HEAL-Link και του διαφορετικού τρόπου λειτουργίας του συστήματος απέναντι σε επισκέπτες και μέλη, οδηγούμαστε στη δημιουργία δύο διαφορετικών μοντέλων εξόρυξης, ένα για κάθε κατηγορία χρηστών. Ο πίνακας JS_WI_WCT_RNS_RCT_V2 περιέχει δύο διαφορετικές ομάδες εγγραφών. Αυτές που καταχωρήθηκαν από τους επισκέπτες και αυτές που καταχωρήθηκαν από τα εγγεγραμμένα μέλη. Τα δύο μοντέλα εξόρυξης που θα δημιουργήσουμε θα αναλύσουν τις δύο αυτές ομάδες δεδομένων ξεχωριστά.

Τα πεδία του πίνακα JS_WI_WCT_RNS_RCT_V2 που διαχωρίζουν τα δεδομένα των επισκεπτών και των μελών, είναι τα πεδία EMAIL και SESSION. Οι τιμές που καταχωρούνται στα πεδία αυτά μας βοηθούν να καταλάβουμε εάν η κάθε εγγραφή καταχωρήθηκε από επισκέπτη ή μέλος.

Όπως έχουμε ήδη αναφέρει, το σύστημα του HEAL-Link χειρίζεται τα πεδία EMAIL και SESSION με τον εξής τρόπο:

- Όταν ο χρήστης είναι επισκέπτης, τότε στο πεδίο EMAIL καταχωρείται η τιμή 'guest' και στο πεδίο SESSION το session id που προσδιορίζει μοναδικά τη συνεδρία του επισκέπτη.
- Όταν ο χρήστης είναι μέλος, τότε στο πεδίο EMAIL καταχωρείται το πραγματικό του e-mail και στο πεδίο SESSION η τιμή 'ok'.

Πίνακας 6.1. Τιμές που καταχωρούνται στα πεδία EMAIL και SESSION του πίνακα JOURNAL_STATS ανάλογα με τον χρήστη

Χρήστης	Τιμή πεδίου EMAIL	Τιμή πεδίου SESSION
Επισκέπτης	'guest'	session_id
Μέλος	Πραγματικό e-mail του χρήστη	'ok'

Με βάση τα παραπάνω, ο ορισμός της συναλλαγής για τους επισκέπτες και τα εγγεγραμμένα μέλη διαφέρει.

Ως συναλλαγή για τους επισκέπτες θεωρείται η στοχοποίηση ψηφιακών συγγραμμάτων κατά τη διάρκεια μίας συνεδρίας. Κάθε συνεδρία ξεκινάει σε μία συγκεκριμένη χρονική στιγμή (όταν ο επισκέπτης εισέλθει στον διαδικτυακό τόπο) και ολοκληρώνεται σε μία δεύτερη συγκεκριμένη χρονική στιγμή (π.χ. όταν κλείσει ο web browser). Αυτές οι δύο χρονικές στιγμές οριοθετούν την έναρξη και τη λήξη της συναλλαγής. Τα αντικείμενα των συναλλαγών είναι τα ψηφιακά συγγράμματα που στοχοποιεί ο χρήστης κατά τη διάρκεια της συνεδρίας.

Το πεδίο που προσδιορίζει μοναδικά τις συναλλαγές των επισκεπτών και που θα χρησιμοποιηθεί ως transaction_id στο μοντέλο παραγωγής κανόνων συσχετίσεων των επισκεπτών, είναι το πεδίο SESSION. Αυτό σημαίνει ότι κάθε session_id είναι και transaction_id.

Για τα δεδομένα των χρηστών μελών, το πεδίο που θα χρησιμοποιηθεί ως transaction_id είναι το πεδίο EMAIL σε συνδυασμό με τα πεδία YEAR, MONTH και DAY. Δηλαδή, ως συναλλαγή για τα εγγεγραμμένα μέλη θεωρούμε

τη στοχοποίηση ψηφιακών συγγραμμάτων κατά τη διάρκεια μίας ημέρας. Κάθε ημέρα είναι και μία ξεχωριστή συναλλαγή.

Με βάση τα παραπάνω, δεν έχουμε παρά να δημιουργήσουμε δύο ξεχωριστά μοντέλα παραγωγής κανόνων συσχετίσεων. Το πρώτο θα αναλύσει τα δεδομένα που παρήγαγαν οι επισκέπτες ενώ το δεύτερο τα δεδομένα των μελών.

2. Τελευταίες ενέργειες προετοιμασίας των δεδομένων

Προτού ξεκινήσουμε την ανάλυση, θα πρέπει να φέρουμε τα δεδομένα του πίνακα JS_WI_WCT_RNS_RCT_V2 σε μορφή κατάλληλη προς εξόρυξη. Δηλαδή, στο λεγόμενο Transactional Layout σύμφωνα με το οποίο τα δεδομένα τροφοδοσίας του αλγόριθμου εξόρυξης πρέπει να αποτελούνται υποχρεωτικά από δύο στήλες (μπορούν να υπάρχουν και περισσότερες στήλες αλλά η ύπαρξη των ακόλουθων δύο στηλών είναι υποχρεωτική):

- Transaction ID
- Item ID

Ήδη προσδιορίσαμε ποιες στήλες θα χρησιμοποιήσουμε ως transaction id ανάλογα με την κατηγορία χρηστών στην οποία ανήκουν τα υπό ανάλυση δεδομένα (επισκεπτών - μελών).

Το Item ID αποτελεί τον κωδικό που προσδιορίζει μοναδικά τα αντικείμενα που περιλαμβάνονται στις συναλλαγές. Ο αλγόριθμος εξόρυξης θα πρέπει να το γνωρίζει ώστε να μπορεί να ξεχωρίζει τα αντικείμενα για να ανακαλύψει τις μεταξύ τους συσχετίσεις. Ως αντικείμενα στην προκειμένη περίπτωση θεωρούνται τα ψηφιακά συγγράμματα. Η στήλη του πίνακα JS_WI_WCT_RNS_RCT_V2 η οποία προσδιορίζει τα συγγράμματα που στοχοποιήθηκαν σε κάθε συναλλαγή, είναι η στήλη JOURNAL.

2.1. Αντικατάσταση τίτλων συγγραμμάτων με τους κωδικούς τους (J_ID)

Ένα πρόβλημα το οποίο θα πρέπει να αντιμετωπίσουμε είναι το γεγονός ότι στη στήλη JOURNAL του πίνακα JS_WI_WCT_RNS_RCT_V2 δεν αποθηκεύονται οι κωδικοί των συγγραμμάτων, έτσι όπως τους έχει καταχωρημένους ο πίνακας JOURNAL (στήλη J_ID), αλλά οι τίτλοι τους. Δεδομένου ότι στην παραγωγή κανόνων συσχετίσεων θα χρησιμοποιήσουμε πληροφορίες ταξινομίας, θα πρέπει να βρούμε ένα τρόπο να αντικαταστήσουμε τους τίτλους των συγγραμμάτων με τους αντίστοιχους κωδικούς τους. Ο λόγος για τον οποίο πρέπει να γίνει αυτό είναι για να μπορούν να αντιστοιχηθούν τα συγγράμματα με τους θεματικούς όρους, τις θεματικές υποκατηγορίες και τις θεματικές κατηγορίες με τις οποίες συσχετίζονται ώστε να είναι δυνατή η χρήση ταξινομίας.

Εάν μας ενδιέφερε να παράγουμε κανόνες που συσχετίζουν μόνο τα συγγράμματα, χωρίς να συμπεριλάβουμε πληροφορίες ταξινομίας, τότε θα μπορούσαμε να χρησιμοποιήσουμε ως item id τους τίτλους των συγγραμμάτων (στήλη JOURNAL του πίνακα JS_WI_WCT_RNS_RCT_V2). Σε αυτή την περίπτωση δε θα χρειαζόταν ούτε αντιστοίχηση ονομάτων.

2.1.1. Πώς θα γίνει η αντικατάσταση

Η αντικατάσταση αυτή θα γίνει με ένα απλό INNER JOIN μεταξύ των πινάκων JS_WI_WCT_RNS_RCT_V2 και JOURNAL. Ο νέος πίνακας που θα δημιουργηθεί θα έχει στη θέση της στήλης JOURNAL τους κωδικούς των συγγραμμάτων (J_ID). Επίσης, θα περιλαμβάνει όλες τις εγγραφές του πίνακα JS_WI_WCT_RNS_RCT_V2 οι οποίες φέρουν στη στήλη JOURNAL ψηφιακό σύγγραμμα το οποίο υπάρχει στον πίνακα JOURNAL. Όλες οι υπόλοιπες εγγραφές (δηλαδή αυτές που φέρουν ψηφιακό σύγγραμμα του οποίου ο τίτλος απουσιάζει από τον πίνακα JOURNAL) δε θα συμπεριληφθούν, καθώς δεν μπορεί να εντοπιστεί κάποιος μοναδικός κωδικός για να αντικαταστήσει τους τίτλους τους.

Είναι λογικό να υπάρχουν τέτοιες εγγραφές, αν αναλογιστούμε ότι σε μεταγενέστερη περίοδο κάποια από τα συγγράμματα που φιλοξενούσε ο κατάλογος και συμμετείχαν σε συναλλαγές, ενδεχομένως να διαγράφηκαν από τον πίνακα JOURNAL. Ένα άλλο ενδεχόμενο είναι απλώς να άλλαξε ο τίτλος τους (με τον κωδικό J_ID να παραμένει ίδιος). Και στην δεύτερη περίπτωση, είναι αδύνατο οι εγγραφές του πίνακα JS_WI_WCT_RNS_RCT_V2 που φέρουν τα συγγράμματα αυτά να αντιστοιχηθούν με τους κωδικούς τους.

2.1.2. Ποιος κωδικός (J_ID) θα χρησιμοποιηθεί στην αντικατάσταση

Αναφέραμε ήδη στο κεφάλαιο κατανόησης των δεδομένων ότι στον πίνακα JOURNAL υπάρχουν 310 μοναδικοί τίτλοι συγγραμμάτων οι οποίοι επαναλαμβάνονται με δύο ή τρεις διαφορετικούς κωδικούς J_ID, είτε γιατί προέρχονται από διαφορετικούς εκδότες είτε γιατί πρόκειται για διαφορετικά συγγράμματα με τον ίδιο τίτλο. Προκύπτει λοιπόν το ερώτημα: Ποιος από τους δύο ή τρεις διαφορετικούς κωδικούς J_ID του ίδιου συγγράμματος θα αντικαταστήσει τον τίτλο του; Θα έχει η επιλογή αυτή επιπτώσεις στους κανόνες που θα παραχθούν με βάση την ταξινομία;

Δεδομένου ότι οι επαναλήψεις των ίδιων συγγραμμάτων (με διαφορετικούς κωδικούς J_ID) ακολουθούν πάντα την ίδια διαδρομή στη θεματική ιεραρχία, όπως αποδείχτηκε στην παράγραφο 2.5 του κεφαλαίου 4, μπορούμε με ασφάλεια να χρησιμοποιήσουμε την εντολή DISTINCT πάνω στο πεδίο TITLE του πίνακα JOURNAL ώστε να αφαιρέσουμε τους επαναλαμβανόμενους τίτλους. Το ποια από τις δύο ή τρεις επαναλήψεις του ίδιου συγγράμματος θα επιλέξει το σύστημα διαχείρισης της βάσης δε θα επηρεάσει τους τελικούς κανόνες, καθώς όλες οι επαναλήψεις του ίδιου συγγράμματος έχουν τον ίδιο τίτλο και ακολουθούν την ίδια διαδρομή στη θεματική ιεραρχία.

3. Δημιουργία μοντέλου επισκεπτών

Η διαδικασία που θα ακολουθήσουμε είναι απλή και συνοψίζεται σε τέσσερα βήματα:

1. **Δημιουργία νέας ροής εξόρυξης (mining flow), σχεδιασμός και υλοποίηση της ροής**

   Η ροή εξόρυξης αποτελεί ένα σχέδιο που περιγράφει αναλυτικά όλες τις ενέργειες που πρέπει να εκτελεστούν για την παραγωγή του μοντέλου.

   Σε μία ροή εξόρυξης έχουμε τη δυνατότητα να ορίσουμε τις πηγές από τις οποίες θα γίνει η άντληση των δεδομένων, να μετατρέψουμε τα δεδομένα φέρνοντάς τα σε μορφή κατάλληλη προς εξόρυξη, να ορίσουμε και να παραμετροποιήσουμε τον αλγόριθμο εξόρυξης καθώς επίσης και τους προορισμούς στους οποίους θα κατευθυνθούν οι κανόνες και τα αποτελέσματα που θα παραχθούν.

2. **Προσθήκη πληροφοριών αναζήτησης ονομάτων στη ροή εξόρυξης**

   Για να είναι κατανοητοί οι κανόνες που θα παράγει το μοντέλο, θα πρέπει να προσθέσουμε στη ροή εξόρυξης πληροφορίες αναζήτησης ονομάτων για την αντικατάσταση των κωδικών των αντικειμένων (κωδικοί συγγραμμάτων, θεματικών ενοτήτων, υποκατηγοριών και κατηγοριών).

3. **Προσθήκη πληροφοριών ταξινομίας στη ροή εξόρυξης**

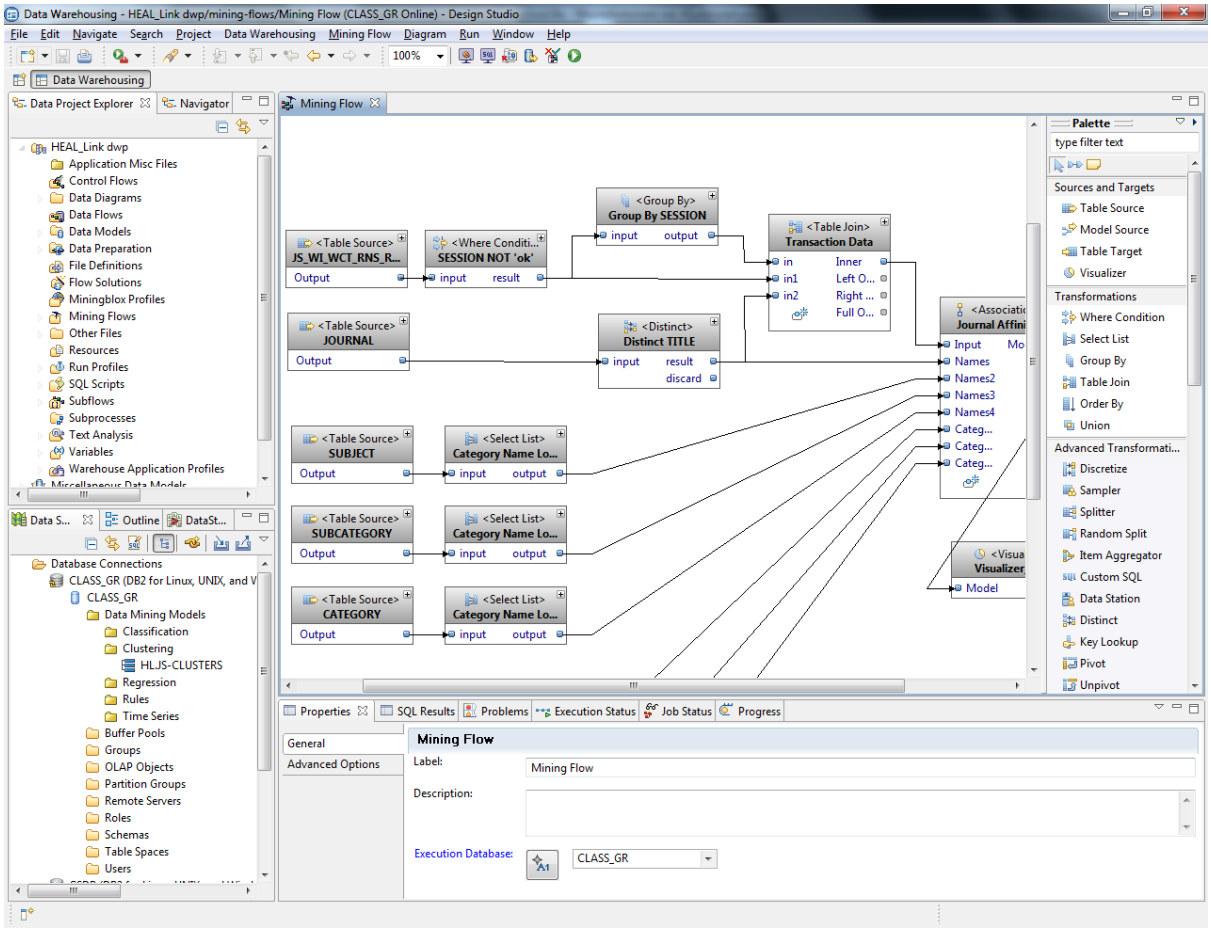
   Για να ανακαλύψουμε εκτός από τις ειδικές συσχετίσεις (σύγγραμμα Α συνεπάγεται σύγγραμμα Β) και πιο γενικές (σύγγραμμα Α συνεπάγεται υποκατηγορία συγγραμμάτων Β), θα πρέπει να προσθέσουμε στη ροή εξόρυξης πληροφορίες ταξινομίας. Ταξινομία είναι μία ιεραρχία κατηγοριών (π.χ. σύγγραμμα - θεματικός όρος - υποκατηγορία θεματικών όρων - κατηγορία θεματικών όρων).

4. **Ανάλυση και αξιολόγηση των αποτελεσμάτων χρησιμοποιώντας τον IM Visualizer**

   Με τον IM Visualizer μπορούμε να μελετήσουμε τους κανόνες που θα παραχθούν και να αξιολογήσουμε το μοντέλο. Ο IM Visualizer παρουσιάζει αναλυτικά τα αποτελέσματα του μοντέλου, δίνοντάς μας τη δυνατότητα να αποφασίσουμε εάν και κατά πόσο ικανοποιεί τις απαιτήσεις που έχουμε θέσει.

Όλα τα παραπάνω θα γίνουν μέσα σε ένα ισχυρό και εύχρηστο περιβάλλον εντός του οποίου ουσιαστικά "ζωγραφίζουμε" τη ροή εξόρυξης και τους διάφορους κόμβους από τους οποίους διέρχονται τα δεδομένα. Αρχικά, υπάρχει ένας άδειος καμβάς πάνω στον οποίο μπορούμε να προσθέσουμε πλήθος χειριστών (operators) για την εξυπηρέτηση διαφορετικών σκοπών: Χειριστές προσδιορισμού πηγών δεδομένων και προορισμών, χειριστές μετατροπής των δεδομένων, χειριστές αλγόριθμων εξόρυξης, χειριστές ανάλυσης και ελέγχου των δεδομένων. Οι χειριστές αυτοί συνδέονται μεταξύ τους με ακμές οι οποίες προσδιορίζουν τη ροή που ακολουθούν τα δεδομένα.

Στο ακόλουθο σχήμα παρουσιάζεται το περιβάλλον εργασίας του Design Studio.



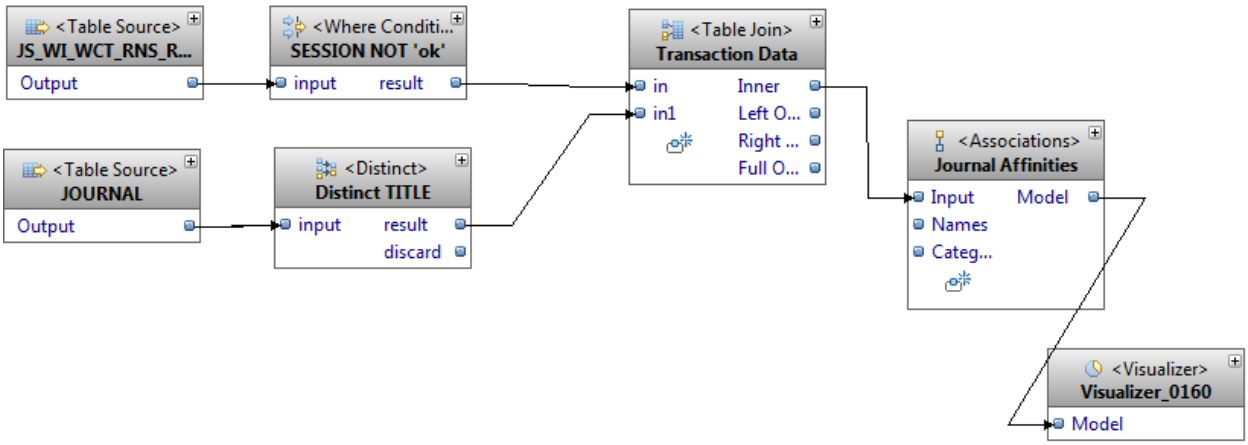
Σχήμα 6.1. Περιβάλλον εργασίας του Design Studio

3.1. Δημιουργία νέας ροής εξόρυξης (mining flow), σχεδιασμός και υλοποίηση της ροής

Στο βήμα αυτό θα εκτελέσουμε τις ακόλουθες έξι ενέργειες:

1. Δημιουργία νέας ροής εξόρυξης
2. Προσθήκη χειριστή συσχετίσεων (associations operator)
3. Προσθήκη χειριστών διαχείρισης πηγών δεδομένων (data source operators)
4. Προσθήκη χειριστή προβολής των αποτελεσμάτων (visualizer operator)
5. Αποθήκευση και δοκιμαστική εκτέλεση της ροής εξόρυξης
6. Προτάσεις βελτιστοποίησης του μοντέλου

Με την ολοκλήρωση των παραπάνω ενεργειών η ροή εξόρυξης θα μοιάζει με τη ροή του ακόλουθου σχήματος:



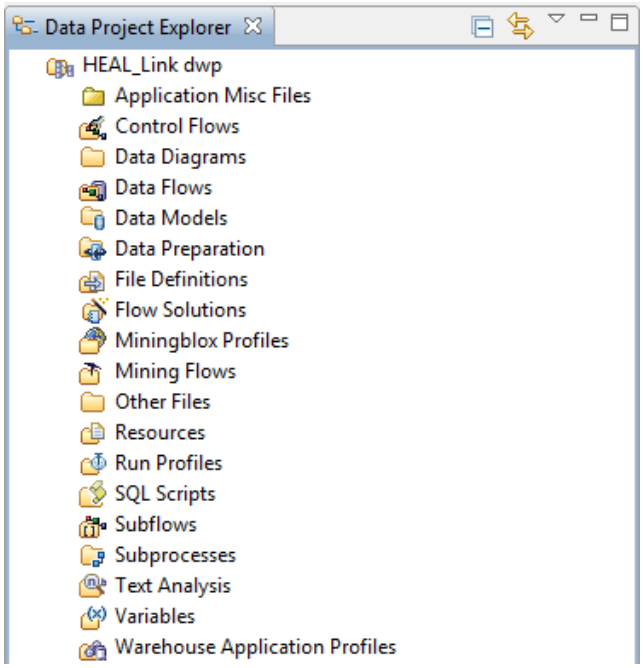
Σχήμα 6.2. Η ροή εξόρυξης για τη δημιουργία του μοντέλου επισκεπτών (1/5)

3.1.1. Δημιουργία νέας ροής εξόρυξης

Πρέπει να δημιουργήσουμε μία νέα κενή ροή εξόρυξης την οποία θα χρησιμοποιήσουμε για να σχεδιάσουμε το μοντέλο εξόρυξης.

Για τη δημιουργία της ροής εξόρυξης:

1. Στον Data Project Explorer επεκτείνουμε το data warehousing project που δημιουργήσαμε (HEAL-Link Data Warehousing Project).



Σχήμα 6.3. Data Project Explorer

2. Δεξί κλικ στον φάκελο **Mining Flows** και επιλογή **New > Mining Flow.**

3. Συμπληρώνουμε τον οδηγό δημιουργίας νέας ροής εξόρυξης με τη βοήθεια του ακόλουθου πίνακα:

Πίνακας 6.2. Πληροφορίες δημιουργίας νέας ροής εξόρυξης

Σελίδα	Βήματα
New Data Mining Flow	• Δίνουμε όνομα στη ροή εξόρυξης:

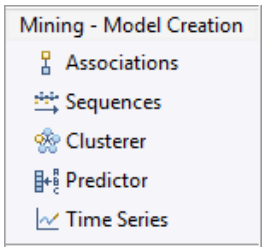
Σελίδα	Βήματα
	`GUEST JOURNAL AFFINITIES` • Επιλέγουμε **Work against database (Online).** Θα εργαστούμε απευθείας στα δεδομένα της βάσης.
Select Connection	Από τη λίστα επιλογής των βάσεων δεδομένων επιλέγουμε **HLDB.**

3.1.2. Προσθήκη χειριστή συσχετίσεων (associations operator)

Ο χειριστής συσχετίσεων είναι υπεύθυνος για την εκτέλεση όλων των λειτουργιών εξόρυξης (ανάλυση δεδομένων, αντιστοίχηση ονομάτων, ταξινομία, παραγωγή κανόνων). Για την προσθήκη και παραμετροποίηση ενός χειριστή συσχετίσεων στη ροή εξόρυξης, κάνουμε τα ακόλουθα:

1. Προσθήκη χειριστή συσχετίσεων

   • Από την παλέτα χειριστών στην δεξιά πλευρά του καμβά, επιλέγουμε τον χειριστή **Associations** (στην ομάδα Data Mining) και τον τοποθετούμε στον καμβά.



Σχήμα 6.4. Ο χειριστής Associations στην παλέτα χειριστών

2. Παραμετροποίηση του χειριστή συσχετίσεων

   • Κάνουμε δεξί κλικ στον χειριστή συσχετίσεων και επιλέγουμε **Show properties view.** Εμφανίζεται το παράθυρο ρυθμίσεων στο κάτω μέρος της οθόνης.

   • Στο πεδίο **Label** της σελίδας General, πληκτρολογούμε Journal Affinities.

   • Στο πεδίο **Model name** της σελίδας Model Name, πληκτρολογούμε GUEST_JOURNAL_AFFINITIES.

   • Στη σελίδα Mining Settings, παραμετροποιούμε το χειριστή συσχετίσεων με βάση τις ρυθμίσεις του ακόλουθου πίνακα:

Πίνακας 6.3. Ρυθμίσεις του χειριστή συσχετίσεων της ροής εξόρυξης GUEST JOURNAL AFFINITIES

Πεδίο	Τιμή	Επεξήγηση
Maximum rule length	3	Το μήκος ενός κανόνα είναι το πλήθος των αντικειμένων της κεφαλής και του σώματος. Θέτοντας το πεδίο maximum rule length στην τιμή 3, το μοντέλο θα μπορεί να παράγει κανόνες που θα περιλαμβάνουν μέχρι τρία αντικείμενα, όπως π.χ. "σύγγραμμα_Α + σύγγραμμα_Β ==> σύγγραμμα_Γ".
Minimum support (%)	0	Επειδή τα προς ανάλυση δεδομένα περιλαμβάνουν πολύ μεγάλο αριθμό εγγραφών (εκατομμύρια), οι κανόνες που θα παραχθούν θα χαρακτηρίζονται από πολύ μικρές τιμές support. Γι' αυτό το λόγο, ορίζουμε ως ελάχιστο support την τιμή 0.

Ολοκληρώσαμε την προσθήκη του χειριστή συσχετίσεων. Στη συνέχεια, θα προσθέσουμε χειριστές που εισάγουν στη ροή εξόρυξης τα δεδομένα.

3.1.3. Προσθήκη χειριστών διαχείρισης πηγών δεδομένων (data source operators)

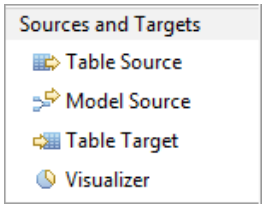
Πρέπει να εισάγουμε στη ροή εξόρυξης χειριστές διαχείρισης πηγών δεδομένων για τη δημιουργία ενός εικονικού πίνακα που θα περιλαμβάνει τα δεδομένα σε μορφή κατάλληλη προς εξόρυξη (Transactional Layout). Ο εικονικός αυτός πίνακας, ο οποίος θα τροφοδοτηθεί στο χειριστή συσχετίσεων, θα αποτελείται από δύο στήλες:

- Transaction ID = **SESSION** (πίνακας JS_WI_WCT_RNS_RCT_V2)
- Item ID = **J_ID** (πίνακας JOURNAL)

Για την εισαγωγή των χειριστών διαχείρισης πηγών δεδομένων και την προεπεξεργασία των δεδομένων, εκτελούμε:

1. Εισαγωγή και ρύθμιση ενός χειριστή Table Source υπεύθυνου να διαχειρίζεται τα δεδομένα του πίνακα JS_WI_WCT_RNS_RCT_V2.

   - Επιλέγουμε από την παλέτα τον χειριστή **Table Source** (ομάδα Sources and Targets) και τον τοποθετούμε στον καμβά.



Σχήμα 6.5. Ο χειριστής Table Source στην παλέτα χειριστών

   - Στο παράθυρο Select Database Table, επιλέγουμε τον πίνακα **JS_WI_WCT_RNS_RCT_V2**.

   Ο πίνακας JS_WI_WCT_RNS_RCT_V2 περιέχει μία εγγραφή για κάθε στοχοποίηση συγγράμματος από ένα χρήστη (επισκέπτη ή μέλος). Οι στήλες που μας ενδιαφέρουν πρωτίστως είναι η στήλη SESSION, που περιέχει τον μοναδικό κωδικό κάθε συνεδρίας και η στήλη JOURNAL που περιέχει τον τίτλο των στοχοποιημένων συγγραμμάτων.

2. Εισαγωγή και ρύθμιση ενός χειριστή Table Source υπεύθυνου να διαχειρίζεται τα δεδομένα του πίνακα JOURNAL.

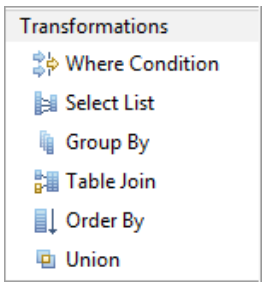
   - Επιλέγουμε από την παλέτα τον χειριστή **Table Source** (ομάδα Sources and Targets) και τον τοποθετούμε στον καμβά.

   - Στο παράθυρο Select Database Table, επιλέγουμε τον πίνακα **JOURNAL**.

   Ο πίνακας JOURNAL περιέχει όλα τα ψηφιακά συγγράμματα τα οποία είναι διαθέσιμα μέσω του διαδικτυακού τόπου HEAL-Link. Οι στήλες που μας ενδιαφέρουν είναι η στήλη J_ID που περιέχει τον μοναδικό κωδικό κάθε συγγράμματος και η στήλη TITLE που περιέχει τον τίτλο κάθε συγγράμματος.

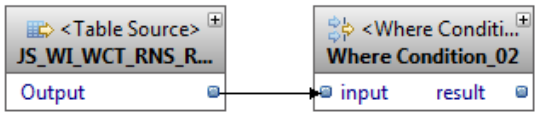
3. Εισαγωγή και ρύθμιση ενός χειριστή <Where Condition> ο οποίος φιλτράρει τα δεδομένα του πίνακα JS_WI_WCT_RNS_RCT_V2 και επιλέγει μόνο τις εγγραφές που καταχωρήθηκαν από επισκέπτες (αποκλείει τις εγγραφές των μελών).

   - Επιλέγουμε από την παλέτα τον χειριστή **<Where Condition>** (ομάδα Transformations) και τον τοποθετούμε στον καμβά.



Σχήμα 6.6. Ο χειριστής Where Condition στην παλέτα χειριστών

- Συνδέουμε τα δεδομένα του πίνακα JS_WI_WCT_RNS_RCT_V2 με το χειριστή <Where Condition>:



Σχήμα 6.7. Σύνδεση χειριστών Table Source και Where Condition

- Κάνουμε κλικ στην πόρτα εξόδου **Output** του χειριστή JS_WI_WCT_RNS_RCT_V2 και στη συνέχεια στην πόρτα εισόδου **input** του χειριστή <Where Condition>.

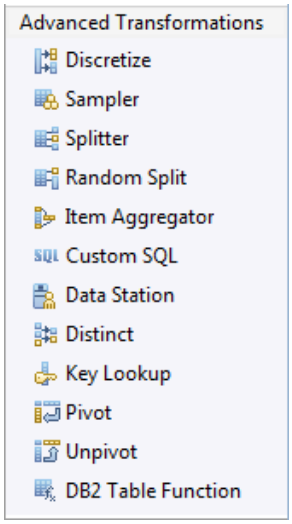
- Κάνοντας διπλό κλικ στο χειριστή <Where Condition> εμφανίζεται ο οδηγός για τη ρύθμισή του. Ακολουθούμε τον οδηγό με βάση τις παρακάτω οδηγίες:

Πίνακας 6.4. Ρυθμίσεις του χειριστή [SESSION NOT 'ok']

Σελίδα	Βήματα
General	Εισάγουμε στο πεδίο Label τον τίτλο: `SESSION NOT 'ok'`
Condition	Εισάγουμε στο πεδίο Filter condition τη συνθήκη: `SESSION NOT LIKE 'ok'` Όλες οι εγγραφές που έχουν 'ok' στο πεδίο SESSION (δηλαδή οι εγγραφές που καταχωρήθηκαν από μέλη) θα αποκλειστούν.

4. Εισαγωγή και ρύθμιση ενός χειριστή <Distinct> ο οποίος επιλέγει από τα δεδομένα του πίνακα JOURNAL μόνο τις εγγραφές με μοναδικό τίτλο.

   - Επιλέγουμε από την παλέτα τον χειριστή **<Distinct>** (ομάδα Advanced Transformations) και τον τοποθετούμε στον καμβά.



Σχήμα 6.8. Ο χειριστής Distinct στην παλέτα χειριστών

- Συνδέουμε τα δεδομένα του πίνακα JOURNAL με το χειριστή <Distinct>.
  - Κάνουμε κλικ στην πόρτα εξόδου **Output** του χειριστή JOURNAL και στη συνέχεια στην πόρτα εισόδου **input** του χειριστή <Distinct>.
- Κάνοντας διπλό κλικ στο χειριστή <Distinct> εμφανίζεται ο οδηγός ρύθμισής του. Ακολουθούμε τον οδηγό με βάση τις παρακάτω οδηγίες:

Πίνακας 6.5. Ρυθμίσεις του χειριστή [Distinct TITLE]

Σελίδα	Βήματα
General	Εισάγουμε στο πεδίο Label τον τίτλο: `Distinct TITLE`
Column Select	Αφαιρούμε από τη λίστα Selected Columns όλες τις στήλες εκτός από τη στήλη TITLE (distinct title).

5. Εισαγωγή και ρύθμιση ενός χειριστή **<Table Join>** ο οποίος συνενώνει τα δεδομένα που παράγουν στις εξόδους τους οι χειριστές [SEESSION Not 'ok'] και [Distinct TITLE]. Συγκεκριμένα, υλοποιεί σύζευξη (inner join) μεταξύ των εγγραφών του πίνακα JS_WI_WCT_RNS_RCT_V2 που καταχωρήθηκαν από επισκέπτες και των μοναδικών συγγραμμάτων του πίνακα JOURNAL, με σκοπό την αντικατάσταση των τίτλων των συγγραμμάτων με τους κωδικούς τους (J_ID).
   - Επιλέγουμε από την παλέτα τον χειριστή **<Table Join>** (ομάδα Transformations) και τον τοποθετούμε στον καμβά.
   - Συνδέουμε τις εξόδους των χειριστών [SESSION Not 'ok'] και [Distinct TITLE] με τον χειριστή <Table Join> ώστε τα δεδομένα που παράγουν οι πρώτοι να τροφοδοτηθούν στον δεύτερο.
     - Κάνουμε κλικ στην πόρτα αποτελεσμάτων **result** του χειριστή [SESSION Not 'ok'] και στη συνέχεια στην πόρτα εισόδου **in** του χειριστή <Table Join>.
     - Κάνουμε κλικ στην πόρτα αποτελεσμάτων **result** του χειριστή [Distinct TITLE] και στη συνέχεια στη δεύτερη πόρτα εισόδου **in1** του χειριστή <Table Join>.
   - Κάνοντας διπλό κλικ στο χειριστή <Table Join> εμφανίζεται ο οδηγός ρύθμισής του. Ακολουθούμε τον οδηγό με βάση τις παρακάτω οδηγίες:

Πίνακας 6.6. Ρυθμίσεις του χειριστή [Transaction Data]

Σελίδα	Βήματα
General	Εισάγουμε στο πεδίο Label τον τίτλο: `Transaction Data`
Condition	Εισάγουμε μία παράσταση SQL η οποία προσδιορίζει ποιες εγγραφές των δύο εικονικών πινάκων στις εισόδους θα συνενωθούν. a. Κλικ στο κουμπί με τα αποσιωπητικά (...). b. Διπλό κλικ στο πεδίο **JOURNAL** του πρώτου πίνακα. c. Διπλό κλικ στον τελεστή ίσον (=). d. Διπλό κλικ στο πεδίο **TITLE** του δεύτερου πίνακα. e. Κλικ στο **OK**. Η συνθήκη σύζευξης θα μοιάζει με την ακόλουθη: `IN_08_0_06_0142.JOURNAL = IN1_08_1_06_0142.TITLE` Οι αριθμοί 08_0_06_0142 και 08_1_06_0142 παράγονται από το Design Studio και προσδιορίζουν τους πίνακες. Ενδέχεται να διαφέρουν από περίπτωση σε περίπτωση. Η παραπάνω συνθήκη ορίζει ότι ο χειριστής [Transaction Data] θα συνδυάσει τις εγγραφές των εικονικών πινάκων [SESSION Not 'ok'] και [Distinct TITLE] που έχουν τους ίδιους τίτλους συγγραμμάτων.
Select List	Ορίζουμε ποιες στήλες θα συμπεριληφθούν στον εικονικό πίνακα που θα δημιουργήσει ο χειριστής [Transaction Data]: a. Κλικ στο εικονίδιο **Delete All** του πίνακα **Result Columns.** b. Διπλό κλικ στο πεδίο **SESSION** του πρώτου πίνακα. c. Διπλό κλικ στο πεδίο **J_ID** του δεύτερου πίνακα.

Ολοκληρώνοντας τον παραπάνω οδηγό, ο χειριστής [Transaction Data] έχει ρυθμιστεί έτσι ώστε να εκπληρώνει δύο σκοπούς:

a. **Αντικαθιστά τους τίτλους των συγγραμμάτων με τους αντίστοιχους κωδικούς τους.**

   Ο λόγος για τον οποίο αντικαθιστούμε τους τίτλους των συγγραμμάτων με τους κωδικούς τους είναι για να μπορούμε να χρησιμοποιήσουμε στο μοντέλο εξόρυξης πληροφορίες ταξινομίας. Να αξιοποιήσουμε δηλαδή τη θεματική ιεραρχία με βάση την οποία είναι ταξινομημένα τα συγγράμματα, με στόχο την παραγωγή γενικότερων αλλά και πλουσιότερων σε γνώση κανόνων. Χρησιμοποιώντας ταξινομία στην εξόρυξη θα προκύψουν κανόνες πιο γενικοί που συσχετίζουν αντικείμενα διαφορετικών επιπέδων (π.χ. συγγράμματα με θεματικές υποκατηγορίες).

   Ο μόνος τρόπος για να γίνει κάτι τέτοιο, είναι να γνωρίζουμε τη διαδρομή που ακολουθεί το κάθε στοχοποιημένο σύγγραμμα στη θεματική ιεραρχία (σύγγραμμα > θεματικός όρος > θεματική υποκατηγορία > θεματική κατηγορία). Μπορούμε να ανακαλύψουμε τη διαδρομή κάθε συγγράμματος στη θεματική ιεραρχία μόνο εάν γνωρίζουμε τον κωδικό του.

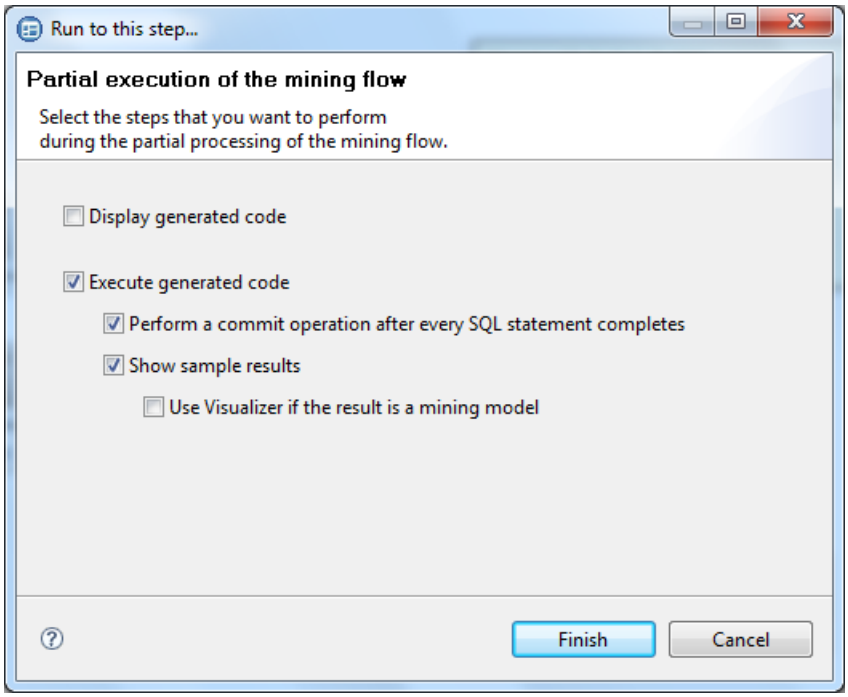
b. **Αφαιρεί από τα τελικά δεδομένα τα πεδία που δε χρειάζονται στον αλγόριθμο εξόρυξης.**

   Η διάταξη δεδομένων συναλλαγών (transactional layout), προϋποθέτει την ύπαρξη δύο πεδίων στα δεδομένα συναλλαγών: Transaction_ID και Item_ID. Προς εκπλήρωση αυτής της προϋπόθεσης, ο χειρι-

στής [Transaction Data] διατηρεί στον τελικό πίνακα δεδομένων συναλλαγών μόνο τις στήλες **SESSION** (transaction id) και **J_ID** (item id).

6. Εκτέλεση της ροής μέχρι αυτό το σημείο για να ελέγξουμε την ορθότητα των προεπεξεργασμένων δεδομένων:

   - Δεξί κλικ στον χειριστή [Transaction Data] και επιλογή **Run to this step.**
   - Στο παράθυρο που εμφανίζεται κάνουμε κλικ στο κουμπί **Finish** αποδεχόμενοι τις προεπιλεγμένες τιμές.



Σχήμα 6.9. Run to this step

Μετά την εκτέλεση της ροής, εμφανίζεται στην ετικέτα Results της προβολής Execution Status ο εικονικός πίνακας που παράγει ο χειριστής [Transaction Data]. Αυτός ο πίνακας θα χρησιμοποιηθεί ως είσοδος στον χειριστή Associations.

SESSION	J_ID
69770CCB3AD972F6E2E900A187941ACB	9277
9622B24D04B71E9F553A33ADD7785669	5456
1B69B4C053638C9E829E14179174A54C	12734
F8906F66F85F6B8F44B470B948801DD1	5452
A0551C601667A708E3EB3DE77693BFC2	5823
78FDB8F34CFEFF14EAD782C01E9580AF	8600

Σχήμα 6.10. Δείγμα δεδομένων του εικονικού πίνακα Transaction Data

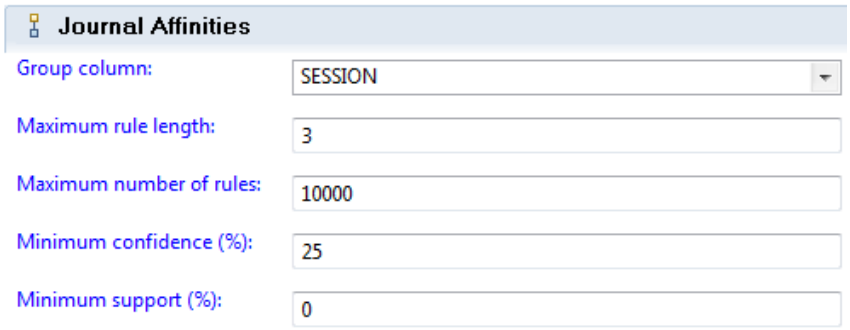
7. Ορισμός του εικονικού πίνακα που παράγει ο χειριστής [Transaction Data] ως είσοδο στον χειριστή [Journal Affinities]:

   - Κάνουμε κλικ στην πόρτα εσωτερικής σύζευξης του χειριστή [Transaction Data] (**Inner**) και στη συνέχεια στην πόρτα εισόδου (**Input**) του χειριστή [Journal Affinities].

8. Ορισμός της στήλης SESSION ως transaction id στον χειριστή [Journal Affinities]:

   - Δεξί κλικ στον χειριστή [Journal Affinities] και επιλογή **Show Properties View.**

- Στη σελίδα Mining Settings, επιλέγουμε στο πεδίο **Group column** την τιμή SESSION. Το πεδίο αυτό ορίζει ότι οι εγγραφές του πίνακα εισόδου που έχουν την ίδια τιμή στο πεδίο SESSION, θα ανήκουν στο ίδιο group (στοιχειοσύνολο). Ορίζουμε δηλαδή ποιο θα είναι το transaction id. Όλα τα συγγράμματα (J_ID) που στοχοποιήθηκαν στην ίδια συνεδρία (SESSION), θα ανήκουν στο ίδιο στοιχειοσύνολο.



Σχήμα 6.11. Mining Settings

Στο σημείο αυτό έχουμε ολοκληρώσει την εκτέλεση των απαραίτητων βημάτων προετοιμασίας των δεδομένων προς ανάλυση. Στη συνέχεια, θα περιγράψουμε τα βήματα που απαιτούνται για την οπτικοποίηση των αποτελεσμάτων που θα παραχθούν από την ανάλυση των δεδομένων.

3.1.4. Προσθήκη χειριστή προβολής των αποτελεσμάτων (visualizer operator)

Για την προβολή του μοντέλου που θα δημιουργήσει η ροή εξόρυξης, θα πρέπει να προσθέσουμε στη ροή τον χειριστή **<Visualizer>**.

1. Προσθήκη του χειριστή <Visualizer> στη ροή εξόρυξης:

   - Επιλέγουμε από την παλέτα τον χειριστή **Visualizer** (ομάδα Sources and Targets) και τον τοποθετούμε στον καμβά.

2. Σύνδεση του χειριστή <Associations> με τον χειριστή <Visualizer>.

   - Κάνουμε κλικ στην πόρτα εξόδου **Model** του χειριστή <Associations> και στη συνέχεια στην πόρτα εισόδου **Model** του χειριστή <Visualizer>.

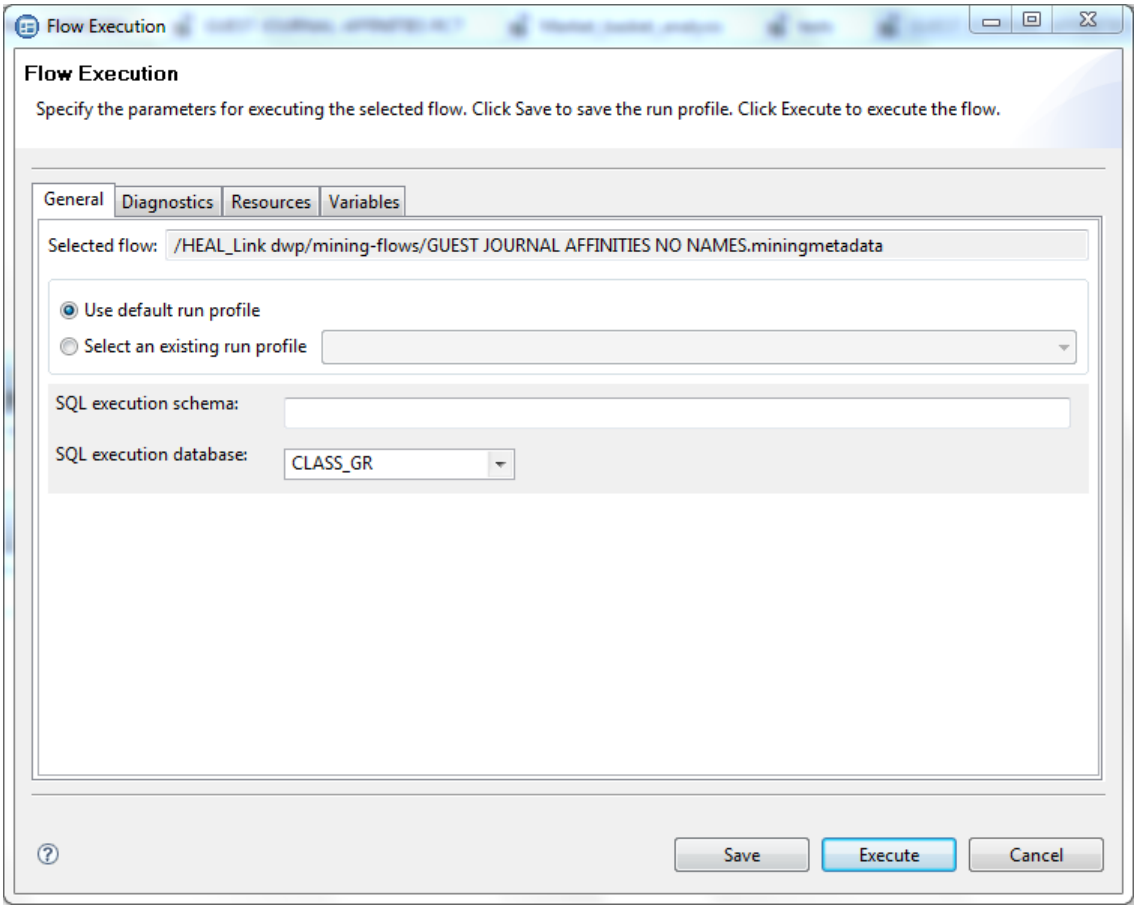
Πλέον, εάν τρέξουμε τη ροή εξόρυξης, το μοντέλο που θα παραχθεί θα ανοίξει απευθείας στον IM Visualizer από όπου θα μπορέσουμε να το μελετήσουμε.

3.1.5. Αποθήκευση και δοκιμαστική εκτέλεση της ροής εξόρυξης

Αποθηκεύουμε τη ροή εξόρυξης και στη συνέχεια την εκτελούμε:

1. Στη μπάρα μενού, επιλέγουμε κατά σειρά: **Mining Flow > Execute.**

2. Στο παράθυρο που ανοίγει (Flow Execution), αποδεχόμαστε τις προεπιλεγμένες τιμές και κάνουμε κλικ στο κουμπί **Execute** για να εκτελεστεί η ροή.



Σχήμα 6.12. Flow Execution

3.1.6. Προβολή του μοντέλου και προτάσεις βελτιστοποίησης

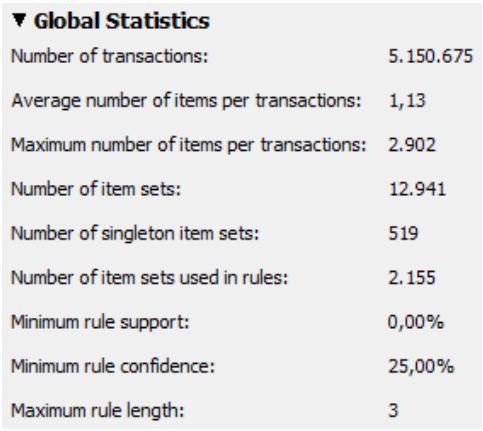
Παρατηρώντας τους κανόνες του μοντέλου στον IM Visualizer (ετικέτα Rules), διαπιστώνουμε ότι είναι δύσκολο να διαβαστούν και να κατανοηθούν, καθώς τα αντικείμενα που περιλαμβάνουν (δηλαδή τα συγγράμματα) αναπαρίστανται με τους κωδικούς τους (J_ID). Για παράδειγμα, ο κανόνας [4851] ==> [4849] δε μας βοηθάει ιδιαίτερα: Χρήστες οι οποίοι επισκέπτονται το σύγγραμμα 4851 επισκέπτονται και το σύγγραμμα 4849.

ID	Rule	▼ Support	Confidence	Lift	Absolute Support
9.472	[4851] ==> [4849]	0,0095%	27,0822%	405,38	491
9.471	[3735] ==> [2045]	0,0081%	25,3325%	727,31	419
9.470	[9771] ==> [5353]	0,0081%	35,0211%	911,02	415
9.469	[571] ==> [2684]	0,0075%	30,3599%	975,51	388
9.468	[6377] ==> [6376]	0,0075%	41,8301%	766,74	384

Σχήμα 6.13. Αναπαράσταση συγγραμμάτων με τους κωδικούς τους

Το μοντέλο μπορεί να βελτιωθεί εάν προσθέσουμε πληροφορίες αναζήτησης ονομάτων (name mapping). Με αυτόν τον τρόπο, ο Intelligent Miner θα γνωρίζει με ποιους τίτλους να αντικαταστήσει τους κωδικούς των συγγραμμάτων. Έτσι, ο παραπάνω κανόνας θα μετατραπεί σε [Medicine and Sciense in Sports and Exercise] ==> [Medicine].

Ένα δεύτερο πρόβλημα το οποίο παρατηρείται στο μοντέλο είναι το μέγιστο πλήθος των αντικειμένων ανά συναλλαγή. Εξετάζοντας τα στατιστικά του μοντέλου (ετικέτα Statistics), διαπιστώνουμε ότι το μέγιστο πλήθος των αντικειμένων ανά συναλλαγή είναι 2.902, αριθμός πολύ μεγάλος για τα ανθρώπινα δεδομένα. Ενώ έχουμε ήδη κάνει μία προσπάθεια εκκαθάρισης των δεδομένων από crawler-sessions, προκύπτει ότι η μέθοδος που ακολουθήσαμε δεν ήταν 100% αποτελεσματική.



▼ Global Statistics

Number of transactions:	5.150.675
Average number of items per transactions:	1,13
Maximum number of items per transactions:	2.902
Number of item sets:	12.941
Number of singleton item sets:	519
Number of item sets used in rules:	2.155
Minimum rule support:	0,00%
Minimum rule confidence:	25,00%
Maximum rule length:	3

Σχήμα 6.14. Πολύ μεγάλο πλήθος αντικειμένων ανά συναλλαγή

Υπενθυμίζεται ότι βασική ιδέα της μεθόδου που ακολουθήσαμε ήταν να θεωρήσουμε ως crawler-sessions εκείνες τις συναλλαγές που περιέχουν περισσότερες από 15 στοχοποιήσεις συγγραμμάτων σε κάποιο από τα λεπτά της διάρκειάς τους.

Για να αντιμετωπίσουμε το πρόβλημα, θα πρέπει να επανέλθουμε στη φάση προετοιμασίας των δεδομένων (η διεργασία εξόρυξης είναι επαναληπτική) και να αφαιρέσουμε (με αποτελεσματικότερο αυτή τη φορά τρόπο) τις συναλλαγές που εκτελέστηκαν από web-crawlers. Θα ομαδοποιήσουμε τις εγγραφές του πίνακα JS_WI_WCT_RNS_RCT_V2 με βάση το πεδίο SESSION και θα αφαιρέσουμε όλες εκείνες τις συναλλαγές που έχουν πλήθος αντικειμένων μεγαλύτερο από 50. Στην τιμή 50 καταλήξαμε μελετώντας τις συναλλαγές που εκτέλεσαν τα εγγεγραμμένα μέλη:

```
select EMAIL, YEAR, MONTH, DAY, count(*) as visits
from db2admin.JS_WI_WCT_RNS_RCT_V2
where session like 'ok'
group by EMAIL, YEAR, MONTH, DAY
order by visits desc
```

Το ερώτημα αυτό μας δείχνει το πλήθος των επισκέψεων των εγγεγραμμένων μελών ανά ημέρα (ταξινομημένα κατά φθίνουσα σειρά).

	EMAIL	YEAR	MONTH	DAY	VISITS
1	[redacted]@balkan.uowm.gr	2008	4	2	95
2	[redacted]@civil.auth.gr	2008	3	10	81
3	[redacted]@jour.auth.gr	2006	2	10	71
4	[redacted]@pspa.uoa.gr	2007	8	29	57
5	[redacted]@jour.auth.gr	2006	1	27	51
6	[redacted]@jour.auth.gr	2006	7	7	50
7	[redacted]@pspa.uoa.gr	2007	2	22	47
8	[redacted]@pp-mail.bio.auth.gr	2007	2	26	46
9	[redacted]@nured.auth.gr	2009	3	3	45
10	[redacted]@jour.auth.gr	2006	3	21	44

Σχήμα 6.15. Επισκέψεις μελών ανά συναλλαγή

Παρατηρούμε ότι το μέγιστο πλήθος των επισκέψεων στις συναλλαγές των εγγεγραμμένων μελών αγγίζει τις 95. Εφόσον λοιπόν τα μέλη μπορούν να κάνουν μέχρι και 95 στοχοποιήσεις σε μία ημέρα (οι οποίες σίγουρα δεν προέρχονται από web-crawlers), μπορούμε να θεωρήσουμε ότι και οι επισκέπτες θα έχουν ανάλογη δυνατότητα.

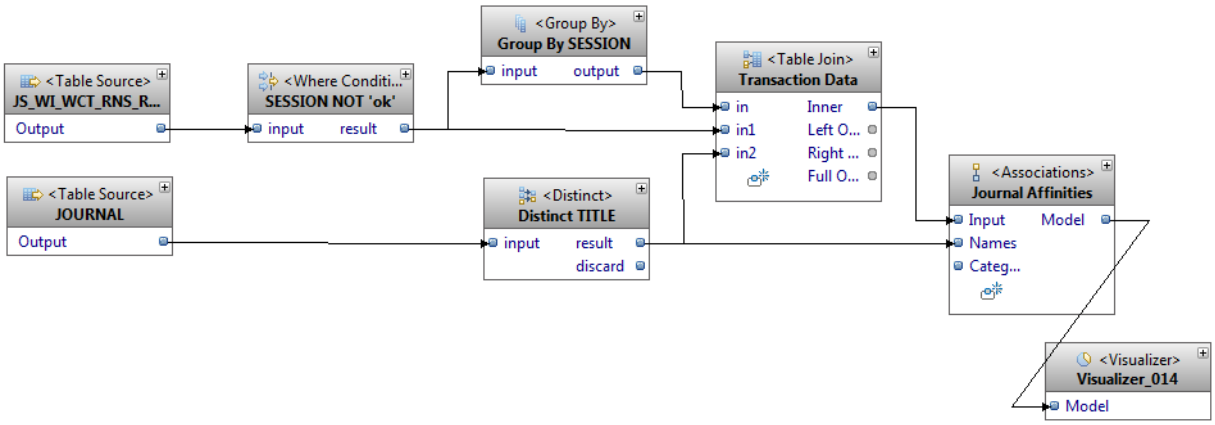
Συνοψίζοντας τις προτάσεις βελτιστοποίησης του μοντέλου, έχουμε:

1. Προσθήκη πληροφοριών αναζήτησης ονομάτων (name mapping).

2. Αφαίρεση των υπολειπόμενων crawler-sessions.

3.2. Προσθήκη στη ροή εξόρυξης πληροφοριών αναζήτησης ονομάτων και αφαίρεση των crawler-sessions

Μετά την ολοκλήρωση των βημάτων βελτιστοποίησης του μοντέλου εξόρυξης, η ροή εξόρυξης θα μοιάζει κάπως έτσι:



Σχήμα 6.16. Η ροή εξόρυξης για τη δημιουργία του μοντέλου επισκεπτών (2/5)

3.2.1. Προσθήκη πληροφοριών αναζήτησης ονομάτων (name mapping)

Για να γίνει η αντικατάσταση των κωδικών (J_ID) των συγγραμμάτων με τους τίτλους τους, θα πρέπει να δείξουμε στον Intelligent Miner πού θα βρει τις πληροφορίες αντιστοίχισης. Οι πληροφορίες αντιστοίχισης βρίσκονται στον εικονικό πίνακα που παράγει ο χειριστής [Distinct TITLE]. Για κάθε σύγγραμμα το οποίο συμμετέχει στην εξόρυξη, φέρει τον κωδικό του και τον τίτλο του.

Η ρύθμιση του χειριστή <Associations> ώστε να χρησιμοποιεί πληροφορίες αναζήτησης ονομάτων, γίνεται ως εξής:

1. Συνδέουμε την έξοδο **result** του χειριστή [Distinct TITLE] στην είσοδο **Names** του [Journal Affinities].

2. Παραμετροποιούμε τον χειριστή [Journal Affinities]:

   - Δεξί κλικ στον χειριστή [Journal Affinities] και επιλογή **Show Properties View.**

   - Στη σελίδα Name Maps, ορίζουμε ποιες στήλες του πίνακα αναζήτησης ονομάτων θα χρησιμοποιήσει ο χειριστής για να αντικαταστήσει τους κωδικούς με τους τίτλους των συγγραμμάτων. Οι επιλογές φαίνονται στον ακόλουθο πίνακα:

Πίνακας 6.7. Ρυθμίσεις αντιστοίχισης ονομάτων

Map Name	Item Id Column	Item Name Column
Names (Το όνομα της πόρτας εισόδου)	J_ID	TITLE

   Με τις παραπάνω ρυθμίσεις θα γνωρίζει ο Intelligent Miner ότι πρέπει να αντικαταστήσει τις τιμές της στήλης J_ID με τις αντίστοιχες τιμές της στήλης TITLE, για κάθε εγγραφή που δέχεται στην πόρτα εισόδου (Input).

   - Στη σελίδα Column Properties, επιλέγουμε την τιμή **Names** στη στήλη **Name Mapping**.

3.2.2. Αφαίρεση των crawler-sessions

Για την αφαίρεση των crawler-sessions, θα ομαδοποιήσουμε τις εγγραφές του εικονικού πίνακα [SESSION Not 'ok'] και θα επιλέξουμε εκείνα τα groups που έχουν πλήθος μικρότερο του 50. Με αυτό τον τρόπο ανακαλύπτουμε τους κωδικούς των συνεδριών κατά τη διάρκεια των οποίων έγιναν λιγότερες από 50 στοχοποιήσεις συγγραμ-

μάτων. Στη συνέχεια, θα συνενώσουμε τις εγγραφές αυτές με τον εικονικό πίνακα [SESSION Not 'ok'] και θα επιλέξουμε τις εγγραφές που έχουν την ίδια τιμή στο πεδίο SESSION. Τέλος, θα πρέπει να γίνει συνένωση των τελευταίων αποτελεσμάτων με τον εικονικό πίνακα [Distinct TITLE] ώστε να αντικατασταθούν οι τίτλοι των συγγραμμάτων με τους κωδικούς τους.

Η υλοποίηση των παραπάνω στη ροή εξόρυξης γίνεται ως εξής:

1. Διαγράφουμε τις προηγούμενες συνδέσεις:
   - Επιλέγουμε τη σύνδεση μεταξύ των χειριστών [SESSION Not 'ok'] και [Transaction Data] και πατάμε DELETE.
   - Επιλέγουμε τη σύνδεση μεταξύ των χειριστών [Distinct TITLE] και [Transaction Data] και πατάμε DELETE.
2. Προετοιμάζουμε τις εισόδους του χειριστή [Transaction Data] για τις νέες συνδέσεις:
   - Κάνουμε δεξί κλικ στην είσοδο **in** του χειριστή [Transaction Data] και επιλέγουμε Edit > Delete UnMapped Columns.
   - Κάνουμε δεξί κλικ στην είσοδο **in1** του χειριστή [Transaction Data] και επιλέγουμε Edit > Delete UnMapped Columns.
3. Προσθέτουμε μία τρίτη είσοδο στον χειριστή [Transaction Data] κάνοντας κλικ στο εικονίδιο που βρίσκεται ακριβώς κάτω από την είσοδο in1.
4. Επιλέγουμε από την παλέτα τον χειριστή **<Group By>** (ομάδα Transformations) και τον τοποθετούμε στον καμβά.
5. Υλοποιούμε τις συνδέσεις μεταξύ των χειριστών:
   - Συνδέουμε την έξοδο **result** του χειριστή [SESSION Not 'ok'] στην είσοδο **input** του χειριστή <Group By>.
   - Συνδέουμε την έξοδο **output** του χειριστή <Group By> στην είσοδο **in** του χειριστή [Transaction Data].
   - Συνδέουμε την έξοδο **result** του χειριστή [SESSION Not 'ok'] στην είσοδο **in1** του χειριστή [Transaction Data].
   - Συνδέουμε την έξοδο **result** του χειριστή [Distinct TITLE] στην είσοδο **in2** του χειριστή [Transaction Data].
6. Κάνοντας διπλό κλικ στο χειριστή <Group By> εμφανίζεται ο οδηγός ρύθμισής του. Ακολουθούμε τον οδηγό με βάση τις παρακάτω οδηγίες:

Πίνακας 6.8. Ρυθμίσεις του χειριστή [Group By SESSION]

Σελίδα	Βήματα
General	Εισάγουμε στο πεδίο Label τον τίτλο: `Group By SESSION`
Select List	Ορίζουμε ποιες στήλες θα συμπεριληφθούν στον εικονικό πίνακα που θα δημιουργήσει ο χειριστής [Group By SESSION]: a. Κλικ στο εικονίδιο **Delete All** του πίνακα **Result Columns.** b. Διπλό κλικ στο πεδίο **SESSION** του πρώτου πίνακα.
Group By	Ορίζουμε πάνω σε ποια στήλη θα γίνει η ομαδοποίηση: a. Κλικ στο εικονίδιο **Delete All** του πίνακα **Result Columns.** b. Διπλό κλικ στο πεδίο **SESSION** του πρώτου πίνακα.
Having	Εισάγουμε στο πεδίο Condition τη συνθήκη με βάση την οποία θα σχηματιστούν τα groups:

Σελίδα	Βήματα
	`COUNT(*) < 50` Η παραπάνω συνθήκη ορίζει ότι στα τελικά αποτελέσματα θα συμμετέχουν μόνο τα groups που έχουν πληθυσμό μικρότερο από 50. Δηλαδή, οι συναλλαγές κατά τη διάρκεια των οποίων έγιναν λιγότερες από 50 στοχοποιήσεις συγγραμμάτων.

7. Ορίζουμε στο χειριστή [Transaction Data] τη συνθήκη με βάση την οποία θα γίνει η συνένωση των τριών εικονικών πινάκων στις εισόδους του και τις στήλες που θα συμπεριληφθούν στην έξοδο.

Πίνακας 6.9. Ρυθμίσεις του χειριστή [Transaction Data]

Σελίδα	Βήματα

Condition	Διαγράφουμε την παλιά συνθήκη και εισάγουμε νέα σύμφωνα με τα ακόλουθα βήματα:
	
	a. Κλικ στο κουμπί με τα αποσιωπητικά (...).
	
	b. Διπλό κλικ στο πεδίο **SESSION** του πρώτου πίνακα.
	
	c. Διπλό κλικ στον τελεστή ίσον (=).
	
	d. Διπλό κλικ στο πεδίο **SESSION** του δεύτερου πίνακα.
	
	e. Διπλό κλικ στο keyword AND.
	
	f. Διπλό κλικ στο πεδίο **JOURNAL** του δεύτερου πίνακα
	
	g. Διπλό κλικ στον τελεστή ίσον (=).
	
	h. Διπλό κλικ στο πεδίο **TITLE** του τρίτου πίνακα.
	
	i. Κλικ στο **OK**.
	
	Η συνθήκη σύζευξης θα μοιάζει με την ακόλουθη:
	
	`IN_0115_0_03.SESSION = IN1_0115_1_03.SESSION`
	`AND IN1_0115_1_03.JOURNAL = IN3_0115_3_03.TITLE`
	
	Η παραπάνω συνθήκη ορίζει ότι ο χειριστής [Transaction Data] θα συνδυάσει τις εγγραφές των εικονικών πινάκων [Group By SESSION], [SESSION Not 'ok'] και [Distinct TITLE] που έχουν τις ίδιες τιμές στα πεδία SESSION, JOURNAL και TITLE.

Select List	Ορίζουμε ποιες στήλες θα συμπεριληφθούν στον εικονικό πίνακα που θα δημιουργήσει ο χειριστής [Transaction Data]: a. Κλικ στο εικονίδιο **Delete All** του πίνακα **Result Columns.** b. Διπλό κλικ στο πεδίο **SESSION** του πρώτου πίνακα. c. Διπλό κλικ στο πεδίο **J_ID** του τρίτου πίνακα.

3.2.3. Αποθήκευση και δοκιμαστική εκτέλεση της ροής εξόρυξης

Αποθηκεύουμε τη ροή εξόρυξης και στη συνέχεια την εκτελούμε:

1. Στη μπάρα μενού, επιλέγουμε κατά σειρά: **Mining Flow > Execute.**

2. Στο παράθυρο που ανοίγει (Flow Execution), αποδεχόμαστε τις προεπιλεγμένες τιμές και κάνουμε κλικ στο κουμπί **Execute** για να εκτελεστεί η ροή.

3.2.4. Προβολή του μοντέλου και προτάσεις βελτιστοποίησης

Παρατηρώντας το μοντέλο μετά τις βελτιωτικές αλλαγές, διαπιστώνουμε ότι και τα δύο προβλήματα που αντιμετωπίσαμε έχουν πλέον λυθεί.

Λόγω της προσθήκης πληροφοριών αναζήτησης ονομάτων, οι κωδικοί των αντικειμένων που συμμετέχουν στους κανόνες έχουν αντικατασταθεί από τους τίτλους των συγγραμμάτων. Πλέον, οι κανόνες μπορούν να μελετηθούν πολύ πιο εύκολα.

ID	Rule	▼ Support	Confidence	Lift	Absolute Support
9.977	[Medicine and Science in Sports and Exercise] ==> [Medicine]	0,0095%	27,0764%	406,10	489
9.976	[Journal of Environmental Education, The] ==> [Environment...	0,0081%	25,3487%	731,40	418
9.975	[Oral Surgery, Oral Medicine, Oral Pathology (continued as O...	0,0080%	34,9662%	910,46	414
9.974	[Archive for History of Exact Sciences] ==> [Historia Mathem...	0,0075%	30,4314%	983,26	388
9.973	[Spine Journal, The] ==> [Spine]	0,0075%	42,0591%	772,81	384

Σχήμα 6.17. Αναπαράσταση συγγραμμάτων με τους τίτλους τους

Επίσης, στην ετικέτα με τα στατιστικά του μοντέλου βλέπουμε ότι το μέγιστο πλήθος αντικειμένων ανά συναλλαγή έχει πέσει στα 45, πράγμα το οποίο σημαίνει ότι στους κανόνες δε συμπεριλαμβάνονται οι ενέργειες των web-crawlers.

▼ Global Statistics	
Number of transactions:	5.150.354
Average number of items per transactions:	1,12
Maximum number of items per transactions:	45
Number of item sets:	12.538
Number of singleton item sets:	549
Number of item sets used in rules:	2.288
Minimum rule support:	0,00%
Minimum rule confidence:	25,00%
Maximum rule length:	3

Σχήμα 6.18. Μέγιστο πλήθος αντικειμένων ανά συναλλαγή χωρίς web-crawlers

Στην προσπάθειά μας να βελτιώσουμε περαιτέρω το μοντέλο, παρατηρούμε τις τιμές support των κανόνων. Οι κανόνες που παρήγαγε το μοντέλο χαρακτηρίζονται από πολύ μικρές τιμές support γεγονός το οποίο οφείλεται στην πολύ σπάνια εμφάνιση των αντικειμένων (συγγραμμάτων) στο πολύ μεγάλο πλήθος συναλλαγών που εξετάστηκαν. Συνολικά εξετάστηκαν 5.150.354 συναλλαγές. Οι επιλογές που κάνουν οι επισκέπτες σε ψηφιακά συγγράμματα καλύπτουν πολύ μεγάλο εύρος τίτλων. Συνδυάζοντας το εύρος των επιλογών με το πλήθος των συναλλαγών, είναι φυσιολογικό οι τιμές του support να παραμένουν σε χαμηλά επίπεδα.

Για τη βελτίωση των κανόνων συσχετίσεων και κυρίως για τη γενίκευση (διεύρυνση) της παραγόμενης γνώσης, μπορούμε να χρησιμοποιήσουμε στο μοντέλο πληροφορίες ταξινομίας.

3.3. Προσθήκη στη ροή εξόρυξης πληροφοριών ταξινομίας

Ταξινομία είναι μία ιεραρχία κατηγοριών με βάση την οποία ταξινομούνται τα αντικείμενα των συναλλαγών. Τα αντικείμενα των συναλλαγών (τα συγγράμματα στην προκειμένη περίπτωση) βρίσκονται στο χαμηλότερο επίπεδο της ιεραρχίας (επίπεδο 0).

Οι κανόνες που έχουμε αυτή τη στιγμή στη διάθεσή μας είναι πολύ ειδικοί, με την έννοια ότι συσχετίζουν μεταξύ τους συγγράμματα τα οποία ενδιαφέρουν σχετικά μικρά ποσοστά του συνόλου των επισκεπτών του διαδικτυακού τόπου. Θέλοντας να γενικεύσουμε την παραγόμενη γνώση και να αυξήσουμε το πλήθος των χρηστών που επηρεάζονται από τους κανόνες, μπορούμε να ανεβούμε ένα επίπεδο πιο πάνω στη θεματική ιεραρχία των συγγραμμάτων (επίπεδο θεματικών όρων) και να συμπεριλάβουμε το επίπεδο αυτό στην εξόρυξη.

Με τον τρόπο αυτό το μοντέλο θα είναι ικανό να παράγει συσχετίσεις όχι μόνο μεταξύ των συγγραμμάτων, [σύγγραμμα A] ==> [σύγγραμμα B], αλλά και μεταξύ θεματικών όρων [θεματικός όρος A] ==> [θεματικός όρος B] και συγγραμμάτων με θεματικούς όρους [σύγγραμμα A] <==> [θεματικός όρος B].

Η κανόνες (και κατ'επέκταση η γνώση) που λαμβάνουμε, θα μπορούσαν να γενικευτούν ακόμα περισσότερο εάν χρησιμοποιούσαμε όλα τα επίπεδα της ιεραρχίας (σύγγραμμα > θεματικός όρος > υποκατηγορία > κατηγορία). Έτσι, θα είχαμε συσχετίσεις μεταξύ συγγραμμάτων και υποκατηγοριών, ή υποκατηγοριών και θεματικών όρων ή κατηγοριών και υποκατηγοριών ή μεταξύ όλων των δυνατών συνδυασμών των τεσσάρων επιπέδων. Είναι σημαντικό να σημειωθεί ότι στους παραγόμενους κανόνες δε θα συμπεριλαμβάνονται μόνο οι γενικοί κανόνες [θεματικός όρος A] <==> [κατηγορία B] αλλά και οι ειδικοί [σύγγραμμα A] ==> [σύγγραμμα B].

Ένα στιγμιότυπο της θεματικής ιεραρχίας των συγγραμμάτων φαίνεται στο ακόλουθο σχήμα:

Σύγγραμμα επίπεδο 0	**Θεματικός όρος** επίπεδο 1	**Υποκατηγορία** επίπεδο 2	**Κατηγορία** επίπεδο 3
Economic Systems Research	Inscriptions -- Arabian Peninsula	History of Asia	History
African Research Bulletin: Political, Social and Cultural Series	Africa -- Social conditions -- 1960 -	History of Africa	
African Studies	Ethnology -- South Africa		

Σχήμα 6.19. Στιγμιότυπο της θεματικής ιεραρχίας

Στην κατηγορία **'History'** ανήκουν οι υποκατηγορίες **'History of Asia'** και **'History of Africa'**. Στις υποκατηγορίες αυτές ανήκουν οι θεματικοί όροι της δεύτερης στήλης. Ο θεματικός όρος **'Inscriptions -- Arabian Peninsula'** ανήκει στην υποκατηγορία **'History of Asia'** και οι θεματικοί όροι **'Africa -- Social conditions -- 1960 -'** και **'Ethnology -- South Africa'** ανήκουν στην υποκατηγορία **'History of Africa'**. Στους θεματικούς όρους της δεύτερης στήλης ανήκουν τα συγγράμματα της πρώτης στήλης.

Στο παρακάτω σχήμα παρουσιάζονται όλοι οι δυνατοί συνδυασμοί συσχετίσεων που μπορεί να παράγει το μοντέλο χρησιμοποιώντας ως ταξινομία ολόκληρη τη θεματική ιεραρχία των συγγραμμάτων.

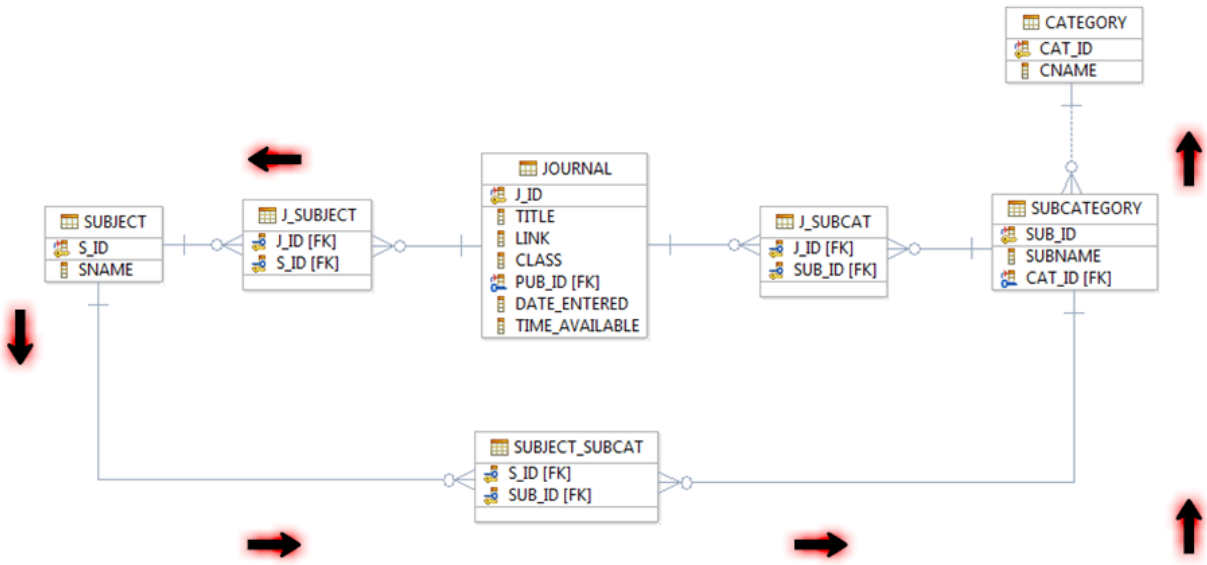
Σ = Σύγγραμμα
ΘΟ = Θεματικός Όρος
Υ = Υποκατηγορία
Κ = Κατηγορία

Σ ==> Σ	Σ <==> ΘΟ	ΘΟ <==> Υ	Υ <==> Κ
ΘΟ ==> ΘΟ	Σ <==> Υ	ΘΟ <==> Κ	
Υ ==> Υ	Σ <==> Κ		
Κ ==> Κ			

Σχήμα 6.20. Πιθανοί συνδυασμοί συσχετίσεων

Η προσθήκη πληροφοριών ταξινομίας θα γίνει βήμα βήμα. Θα ξεκινήσουμε προσθέτοντας το πρώτο επίπεδο της θεματικής ιεραρχίας (τους θεματικούς όρους) και στη συνέχεια θα προσθέσουμε τις υποκατηγορίες και τις κατηγορίες. Με την ολοκλήρωση κάθε βήματος, θα εκτελούμε τη ροή εξόρυξης και θα ελέγχουμε τα αποτελέσματα.

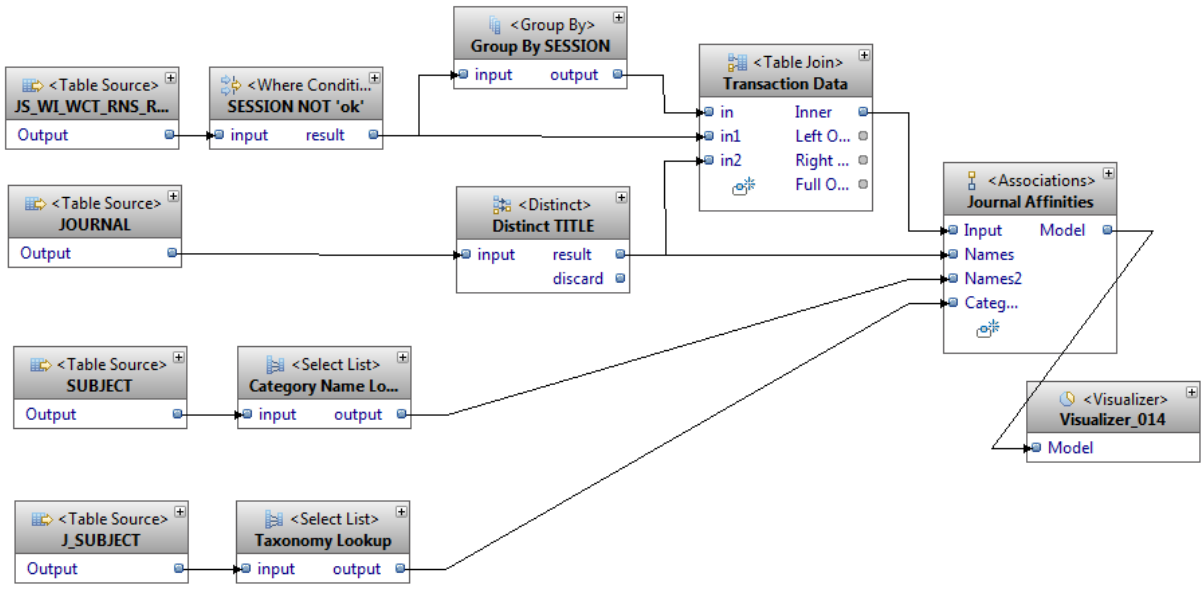
Οι πίνακες που θα χρησιμοποιήσουμε για να προσθέσουμε πληροφορίες ταξινομίας στο μοντέλο εξόρυξης υποδεικνύονται από τα βελάκια του ακόλουθου σχήματος:



Σχήμα 6.21. Οι πίνακες που θα χρησιμοποιηθούν για την προσθήκη πληροφοριών ταξινομίας

3.3.1. Προσθήκη επιπέδου θεματικών όρων

Με την ολοκλήρωση της προσθήκης του επιπέδου θεματικών όρων στις πληροφορίες ταξινομίας του μοντέλου, η ροή εξόρυξης θα μοιάζει με τη ροή του ακόλουθου σχήματος:



Σχήμα 6.22. Η ροή εξόρυξης για τη δημιουργία του μοντέλου επισκεπτών (3/5)

Για να προσθέσουμε στη ροή εξόρυξης το πρώτο επίπεδο της θεματικής ιεραρχίας, τους θεματικούς όρους, θα πρέπει να προσθέσουμε ένα νέο χειριστή πηγών δεδομένων ο οποίος θα εισάγει στο μοντέλο τα δεδομένα του πίνακα J_SUBJECT. Ο πίνακας J_SUBJECT συσχετίζει τα συγγράμματα με τους θεματικούς όρους. Αποτελεί δηλαδή τον πίνακα ταξινομίας ή αλλιώς τον χάρτη κατηγοριών.

Πίνακας 6.10. Οι στήλες παιδιού και γονέα του πίνακα ταξινομίας J_SUBJECT

Παιδί	Γονέας
J_ID	S_ID

Επίσης, θα εισάγουμε έναν πίνακα αναζήτησης ονομάτων για την αντικατάσταση των κωδικών των θεματικών όρων με τις αντίστοιχες ονομασίες τους έτσι ώστε οι τελευταίοι να εμφανίζονται στους παραγόμενους κανόνες

με τις ονομασίες και όχι με τους κωδικούς τους. Ο πίνακας αυτός είναι ο πίνακας SUBJECT. Για κάθε θεματικό όρο, φέρει τον κωδικό και την ονομασία του.

Ο Intelligent Miner θέτει ένα βασικό περιορισμό που αφορά τον πίνακα ταξινομίας: Η ταξινομία πρέπει να είναι ακυκλική. Μία κατηγορία δεν επιτρέπεται να είναι ούτε άμεσο ούτε έμμεσο μέλος του εαυτού της.

Αυτό σημαίνει ότι δεν πρέπει να υπάρχει στον πίνακα ταξινομίας καμία τιμή που να είναι κοινή στα πεδία γονέα και παιδιού. Για παράδειγμα, εάν σε κάποια εγγραφή του πίνακα ταξινομίας J_SUBJECT υπάρχει η τιμή 4 στο πεδίο J_ID, τότε η ίδια τιμή (4) δεν πρέπει να εμφανίζεται πουθενά στη στήλη S_ID.

Εάν δεν τηρηθεί ο παραπάνω περιορισμός, τότε ο Intelligent Miner θα διακόψει την εκτέλεση της ροής επιστρέφοντας το ακόλουθο μήνυμα σφάλματος:

> **3515 There are too many taxonomy levels, or the taxonomy is cyclic.**
>
> **Explanation:** Intelligent Miner does not accept more than 255 taxonomy levels or cyclic taxonomies.

Εξετάζοντας τα δεδομένα του πίνακα J_SUBJECT διαπιστώνουμε ότι ο περιορισμός που θέτει ο Intelligent Miner παραβιάζεται. Για παράδειγμα, υπάρχει κωδικός J_ID με τιμή 1 και κωδικός S_ID με τιμή 1. Για να αντιμετωπίσουμε το πρόβλημα αυτό, θα επιλέξουμε έναν αριθμό τον οποίο θα προσθέσουμε σε όλες τις τιμές S_ID του πίνακα J_SUBJECT. Ο αριθμός αυτός θα πρέπει να είναι μεγαλύτερος από τη μέγιστη τιμή της στήλης J_ID έτσι ώστε να είμαστε βέβαιοι ότι κανένα από τα αθροίσματα που θα προκύψουν δε θα υπάρχει στη στήλη J_ID. Ο αριθμός τον οποίο επιλέγουμε είναι το 10.000.000 (εξηγούμε το γιατί παρακάτω).

Πίνακας 6.11. Προσθήκη της τιμής 10.000.000 στους κωδικούς S_ID

J_ID	S_ID
1	14.441 (+10.000.000)
2	2706 (+10.000.000)
3	13 (+10.000.000)
3	22 (+10.000.000)
4	1 (+10.000.000)
4	2 (+10.000.000)

Η αλλαγή στις τιμές του πεδίου S_ID θα γίνει σε εικονικό πίνακα και όχι στη βάση δεδομένων. Θα εισάγουμε στη ροή εξόρυξης ένα χειριστή <Select List> στον οποίο θα συνδέσουμε τον πίνακα J_SUBJECT. Ο χειριστής <Select List> θα παράγει στην έξοδό του έναν εικονικό πίνακα με δύο στήλες: J_ID και S_ID, με όλες τις τιμές του τελευταίου να είναι αυξημένες κατά 10.000.000.

Δεδομένης της αύξησης των τιμών της στήλης S_ID του πίνακα J_SUBJECT, η ίδια ακριβώς αλλαγή θα πρέπει να γίνει και στη στήλη S_ID του πίνακα SUBJECT ώστε να μη χαλάσει η αντιστοίχιση των ονομάτων με τους κωδικούς των θεματικών όρων. Για αυτό το λόγο, θα χρησιμοποιήσουμε ένα δεύτερο χειριστή <Select List> ο οποίος θα τροποποιήσει ανάλογα τις τιμές του πεδίου S_ID του πίνακα SUBJECT.

Ακολουθούν τα βήματα προσθήκης πληροφοριών ταξινομίας στη ροή εξόρυξης και παραμετροποίησης του χειριστή συσχετίσεων [Journal Affinities].

3.3.1.1. Εισαγωγή των δεδομένων ταξινομίας και αναζήτησης ονομάτων

1. Εισαγωγή και ρύθμιση ενός χειριστή Table Source υπεύθυνου να διαχειρίζεται τα δεδομένα του πίνακα J_SUBJECT.
   - Επιλέγουμε από την παλέτα τον χειριστή **Table Source** (ομάδα Sources and Targets) και τον τοποθετούμε στον καμβά.
   - Στο παράθυρο Select Database Table, επιλέγουμε τον πίνακα **J_SUBJECT**.
2. Εισαγωγή και ρύθμιση ενός χειριστή <Select List> υπεύθυνου για τη δημιουργία εικονικού πίνακα από τα δεδομένα του πίνακα J_SUBJECT, με προσαυξημένες τις τιμές του πεδίου S_ID.

- Επιλέγουμε από την παλέτα τον χειριστή **Select List** (ομάδα Transformations) και τον τοποθετούμε στον καμβά.
- Σύνδεση του χειριστή [J_SUBJECT] με τον χειριστή <Select List>.
  - Κάνουμε κλικ στην πόρτα εξόδου του χειριστή [J_SUBJECT] (Output) και στη συνέχεια στην πόρτα εισόδου του χειριστή <Select List> (input).
- Παραμετροποίηση του χειριστή <Select List>:
  - Δεξί κλικ στον χειριστή <Select List> και επιλογή **Show Properties View.**
  - Ρυθμίζουμε το χειριστή <Select List> με βάση τις οδηγίες του παρακάτω πίνακα:

Πίνακας 6.12. Ρυθμίσεις του χειριστή [Taxonomy Lookup] (J_SUBJECT)

Σελίδα	Βήματα
General	Εισάγουμε στο πεδίο Label τον τίτλο: `Taxonomy Lookup`
Select List	Ορίζουμε ποιες στήλες θα συμπεριληφθούν στον εικονικό πίνακα που θα δημιουργήσει ο χειριστής [Taxonomy Lookup]. Διατηρούμε τις εγγραφές που υπάρχουν ήδη στον πίνακα Result Columns (J_ID, S_ID). Για να προσθέσουμε τον αριθμό 10.000.000 σε όλες τις τιμές S_ID του εικονικού πίνακα, κάνουμε τα ακόλουθα: a. Κλικ στην τιμή S_ID της στήλης Expression στον πίνακα Result Columns. b. Κλικ στο κουμπί με τα αποσιωπητικά που εμφανίζεται στη δεξιά άκρη του κελιού. c. Στον SQL Expression Builder πηγαίνουμε στο πεδίο SQL Text και προσθέτουμε στην ήδη υπάρχουσα τιμή το "+ 10.000.000". Η έκφραση θα πρέπει να μοιάζει με την ακόλουθη: `"INPUT_024_0"."S_ID" + 10000000` d. Πατάμε OK.

3. Εισαγωγή και ρύθμιση ενός χειριστή Table Source υπεύθυνου να διαχειρίζεται τα δεδομένα του πίνακα SUBJECT.
   - Επιλέγουμε από την παλέτα τον χειριστή **Table Source** (ομάδα Sources and Targets) και τον τοποθετούμε στον καμβά.
   - Στο παράθυρο Select Database Table, επιλέγουμε τον πίνακα **SUBJECT**.
4. Εισαγωγή και ρύθμιση ενός χειριστή <Select List> υπεύθυνου για τη δημιουργία ενός εικονικού πίνακα από τα δεδομένα του πίνακα SUBJECT, με προσαυξημένες τις τιμές του πεδίου S_ID.
   - Επιλέγουμε από την παλέτα τον χειριστή **Select List** (ομάδα Transformations) και τον τοποθετούμε στον καμβά.
   - Σύνδεση του χειριστή [SUBJECT] με τον χειριστή <Select List>.
     - Κάνουμε κλικ στην πόρτα εξόδου του χειριστή [SUBJECT] (Output) και στη συνέχεια στην πόρτα εισόδου του χειριστή <Select List> (input).
   - Παραμετροποίηση του χειριστή <Select List>:
     - Δεξί κλικ στον χειριστή <Select List> και επιλογή **Show Properties View.**

- Ρυθμίζουμε το χειριστή <Select List> με βάση τις οδηγίες του παρακάτω πίνακα:

Πίνακας 6.13. Ρυθμίσεις του χειριστή [Category Name Lookup] (SUBJECT)

Σελίδα	Βήματα
General	Εισάγουμε στο πεδίο Label τον τίτλο: `Category Name Lookup`

Select List	Ορίζουμε ποιες στήλες θα συμπεριληφθούν στον εικονικό πίνακα που θα δημιουργήσει ο χειριστής [Category Name Lookup]. Διατηρούμε τις εγγραφές που υπάρχουν ήδη στον πίνακα Result Columns (S_ID, SNAME). Για να προσθέσουμε τον αριθμό 10.000.000 σε όλες τις τιμές S_ID του εικονικού πίνακα, κάνουμε τα ακόλουθα:
	a. Κλικ στην τιμή S_ID της στήλης Expression στον πίνακα Result Columns.
	b. Κλικ στο κουμπί με τα αποσιωπητικά που εμφανίζεται στη δεξιά άκρη του κελιού.
	c. Στον SQL Expression Builder πηγαίνουμε στο πεδίο SQL Text και προσθέτουμε στην ήδη υπάρχουσα τιμή το "+ 10.000.000". Η έκφραση θα πρέπει να μοιάζει με την ακόλουθη:
	`"INPUT_024_0"."S_ID" + 10000000`
	d. Πατάμε OK.
	Στο επόμενο βήμα θα προσθέσουμε μπροστά από τις ονομασίες των θεματικών όρων το πρόθεμα 'Subj', έτσι ώστε να μπορούμε να διακρίνουμε στους κανόνες συσχετίσεων τα συγγράμματα και τους θεματικούς όρους.
	a. Κλικ στην τιμή SNAME της στήλης Expression στον πίνακα Result Columns.
	b. Κλικ στο κουμπί με τα αποσιωπητικά που εμφανίζεται στη δεξιά άκρη του κελιού.
	c. Στον SQL Expression Builder πηγαίνουμε στο πεδίο SQL Text και προσθέτουμε στην αρχή της ήδη υπάρχουσας τιμής το ['Subj: ' ‖] (χωρίς τις αγκύλες). Η έκφραση θα πρέπει να μοιάζει με την ακόλουθη:
	`'Subj: ' || "INPUT_023_0"."SNAME"`
	d. Πατάμε OK.

Η εισαγωγή των δεδομένων ταξινομίας και αναζήτησης ονομάτων για την κατηγορία θεματικών όρων ολοκληρώθηκε. Στη συνέχεια θα ρυθμίσουμε τον χειριστή συσχετίσεων ώστε να γνωρίζει που βρίσκονται αλλά και πως θα χρησιμοποιήσει τα δεδομένα αυτά.

3.3.1.2. Παραμετροποίηση του χειριστή συσχετίσεων [Journal Affinities]

1. Προσθήκη νέας πόρτας εισόδου ονομάτων (Names2) στον χειριστή [Journal Affinities] (ενδέχεται να ονομάζεται Names1).

   - Κάνουμε κλικ στο εικονίδιο που βρίσκεται ακριβώς κάτω από την πόρτα εισόδου Category. Στο παράθυρο που εμφανίζεται επιλέγουμε Names και πατάμε OK. Μία νέα πόρτα (Names2) προστέθηκε στον χειριστή [Journal Affinities].

2. Σύνδεση των χειριστών [Taxonomy Lookup] και [Category Name Lookup] με τον χειριστή [Journal Affinities].

- Κάνουμε κλικ στην πόρτα εξόδου **output** του χειριστή [Taxonomy Lookup] και στη συνέχεια στην πόρτα εισόδου **Category** του χειριστή [Journal Affinities].

- Κάνουμε κλικ στην πόρτα εξόδου **output** του χειριστή [Category Name Lookup] και στη συνέχεια στην πόρτα εισόδου **Names2** του χειριστή [Journal Affinities].

3. Ρύθμιση του χειριστή [Journal Affinities] ώστε να χρησιμοποιήσει τις πληροφορίες ταξινομίας.

   - Κάνουμε διπλό κλικ στον χειριστή [Journal Affinities] και συμπληρώνουμε τον οδηγό με βάση τα ακόλουθα βήματα:

Πίνακας 6.14. Ρυθμίσεις του χειριστή [Journal Affinities] ώστε να χρησιμοποιεί ταξινομία

Σελίδα	Βήματα

Name Maps	Ορίζουμε τις ακόλουθες τιμές στα πεδία του χάρτη αντιστοίχισης ονομάτων Names2:
	
	Item ID Column
	
	`S_ID`
	
	Item Name
	
	`SNAME`
	
	Τα παραπάνω πεδία καθορίζουν ποια στήλη του πίνακα αντιστοίχισης ονομάτων περιλαμβάνει τους κωδικούς και ποια τα ονόματα που θα αντικαταστήσουν τους κωδικούς των θεματικών όρων.

Taxonomy	Στα πεδία του πίνακα Category Map Definition, ορίζουμε τις ακόλουθες τιμές:
	
	Child Column
	
	`J_ID`
	
	Parent Column
	
	`S_ID`
	
	Recursive Map
	
	`No`
	
	Name Mapping
	
	`Names2`
	
	Οι τιμές των πεδίων **Child Column** και **Parent Column** ορίζουν ότι τα αντικείμενα που προσδιορίζονται από τη στήλη J_ID στα δεδομένα συναλλαγών (δηλαδή τα συγγράμματα) είναι μέλη των κατηγοριών που προσδιορίζονται από τη στήλη S_ID στον πίνακα ταξινομίας (δηλαδή των θεματικών όρων).
	
	Η τιμή No στο πεδίο **Recursive Map** ορίζει ότι ένα αντικείμενο δεν μπορεί να είναι ταυτόχρονα γονέας και παιδί.
	
	Το πεδίο **Name Mapping** ορίζει ότι ο πίνακας αντιστοίχισης ονομάτων του πίνακα ταξινομίας είναι συνδεδεμένος στην πόρτα εισόδου Names2 του χειριστή [Journal Affinities].

Column Properties	Στο πεδίο Taxonomy της εισερχόμενης στήλης J_ID ορίζουμε την ακόλουθη τιμή: `Yes`

Σελίδα	Βήματα
	Με τη συγκεκριμένη ρύθμιση γνωστοποιούμε στον χειριστή [Journal Affinities] ότι υπάρχουν διαθέσιμες πληροφορίες ταξινομίας για τα αντικείμενα της στήλης J_ID του πίνακα δεδομένων συναλλαγών.

Στο σημείο αυτό η ροή εξόρυξης είναι ρυθμισμένη κατάλληλα ώστε να δημιουργήσει ένα μοντέλο το οποίο θα περιλαμβάνει κανόνες που συσχετίζουν αντικείμενα δύο διαφορετικών κατηγοριών (συγγράμματα και θεματικούς όρους).

3.3.1.3. Εκτέλεση της ροής εξόρυξης και προβολή του μοντέλου

Αφού εκτελέσουμε τη ροή εξόρυξης, το μοντέλο που παράγεται ανοίγει απευθείας στον IM Visualizer. Ένα δείγμα των κανόνων φαίνεται στο ακόλουθο σχήμα:

Rule	▼ Support	Confidence	Lift	Absolute Support
[Subj: Heart -- Diseases] ==> [Subj: Cardiology]	0,0748%	33,1954%	61,26	3.852
[Subj: Food -- Composition] ==> [Subj: Food -- Analysis]	0,0592%	39,4873%	118,87	3.050
[Subj: Food -- Composition] ==> [Subj: Food]	0,0546%	36,4190%	137,07	2.813
[Subj: Food -- Composition]+[Subj: Food -- Analysis] ==> [Subj: Food]	0,0546%	92,2295%	347,13	2.813
[Subj: Food]+[Subj: Food -- Analysis] ==> [Subj: Food -- Composition]	0,0546%	84,0705%	560,58	2.813
[Subj: Food]+[Subj: Food -- Composition] ==> [Subj: Food -- Analysis]	0,0546%	100,0000%	301,03	2.813
[Subj: Agricultural chemistry] ==> [Subj: Plants -- Nutrition]	0,0479%	49,7980%	368,34	2.465
[Subj: Plants -- Nutrition] ==> [Subj: Agricultural chemistry]	0,0479%	35,4014%	368,34	2.465
[Subj: Plants -- Nutrition] ==> [Subj: Food -- Analysis]	0,0475%	35,1285%	105,75	2.446
[Subj: Agricultural chemistry] ==> [Subj: Food -- Analysis]	0,0475%	49,3737%	148,63	2.444

Σχήμα 6.23. Δείγμα κανόνων συσχετίσεων μεταξύ θεματικών όρων

Στους κανόνες φαίνεται ξεκάθαρα το πρόθεμα 'Subj' το οποίο ξεχωρίζει τα αντικείμενα που αποτελούν τους θεματικούς όρους από τα αντικείμενα που αποτελούν τα συγγράμματα (χωρίς πρόθεμα). Ας επιχειρήσουμε να διαβάσουμε τον πρώτο κανόνα:

[Subj: Heart -- Diseases] ==> [Subj: Cardiology]

Χρήστες οι οποίοι στοχοποιούν συγγράμματα που είναι συσχετισμένα με τον θεματικό όρο **Heart -- Diseases**, στοχοποιούν επίσης συγγράμματα που είναι συσχετισμένα με τον θεματικό όρο **Cardiology**.

Παρατηρώντας τις τιμές support των κανόνων, διαπιστώνουμε ότι έχουν αυξηθεί σε σχέση με τους κανόνες του προηγούμενου μοντέλου. Επίσης, οι κανόνες με τις υψηλότερες τιμές support φαίνεται να αποτελούνται μόνο από θεματικούς όρους (πρόκειται για κανόνες της μορφής [θεματικός όρος A] ==> [θεματικός όρος B]). Βεβαίως, πιο κάτω στο μοντέλο υπάρχουν κανόνες που συσχετίζουν θεματικούς όρους με συγγράμματα (η ταξινόμηση είναι φθίνουσα με βάση το support) οι οποίοι όμως έχουν πολύ μικρότερες τιμές υποστήριξης. Μερικοί από αυτούς φαίνονται στο ακόλουθο σχήμα:

Rule	▼ Support	Confidence	Lift	Absolute Support
[Historia Mathematica] ==> [Subj: Science -- History]	0,0115%	37,1393%	382,03	592
[Food Research International] ==> [Subj: Food -- Analysis]	0,0107%	35,5943%	107,15	551
[Medicine and Science in Sports and Exercise] ==> [Subj: Medicine]	0,0098%	27,9070%	27,86	504
[Medicine and Science in Sports and Exercise] ==> [Medicine]	0,0095%	27,0764%	406,10	489
[Food Science and Technology International] ==> [Subj: Food]	0,0089%	37,2661%	140,26	458
[Physiologia Plantarum] ==> [Subj: Botany]	0,0084%	37,6950%	115,27	435
[Journal of Environmental Education, The] ==> [Environmental Educ...	0,0081%	25,3487%	731,40	418
[Oral Surgery, Oral Medicine, Oral Pathology (continued as Oral Surg...	0,0080%	34,9662%	910,46	414
[Food Control] ==> [Subj: Food]	0,0078%	29,6841%	111,72	404
[Journal of Polymer Science Part A: Polymer Chemistry] ==> [Subj: P...	0,0077%	28,1740%	91,97	395

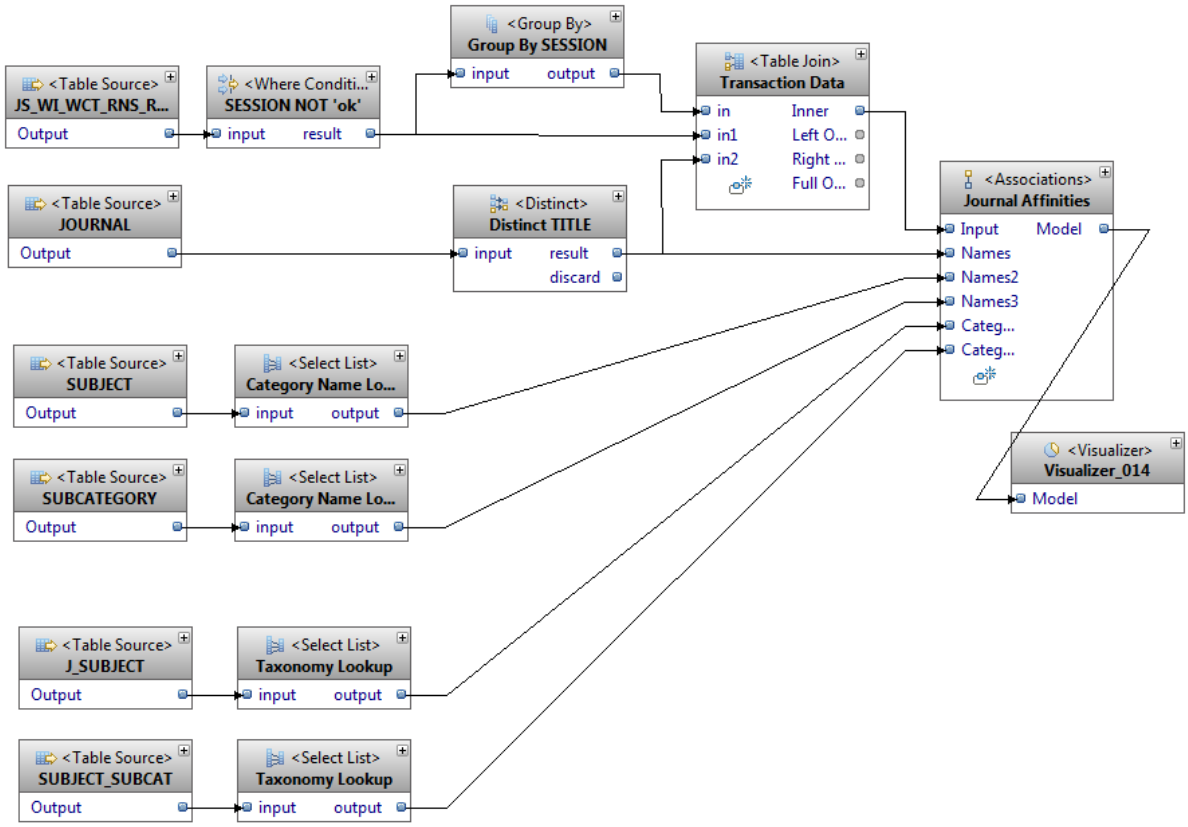
Σχήμα 6.24. Μερικοί κανόνες που συσχετίζουν ψηφιακά συγγράμματα με θεματικούς όρους

Σίγουρα οι κανόνες που περιλαμβάνουν και συγγράμματα είναι πολύ περισσότεροι από όσους εμφανίζει το μοντέλο. Λόγω όμως του περιορισμού που θέσαμε στον αλγόριθμο εξόρυξης (το πολύ 10.000 κανόνες) και του γε-

γονότος ότι ο Intelligent Miner περιλαμβάνει στα μοντέλα τους κανόνες με τις καλύτερες τιμές support, πολλοί κανόνες που συσχετίζουν συγγράμματα μεταξύ τους και συγγράμματα με θεματικούς όρους, οι οποίοι έχουν πολύ χαμηλές τιμές support, μένουν εκτός μοντέλου. Εάν αυξήσουμε το μέγιστο πλήθος κανόνων που μπορεί να επιστρέψει το μοντέλο, τότε θα αποκαλύψουμε πολλούς από αυτούς.

3.3.2. Προσθήκη επιπέδου θεματικών υποκατηγοριών

Με την ολοκλήρωση της προσθήκης του επιπέδου θεματικών υποκατηγοριών στις πληροφορίες ταξινομίας του μοντέλου, η ροή εξόρυξης θα μοιάζει κάπως έτσι:



Σχήμα 6.25. Η ροή εξόρυξης για τη δημιουργία του μοντέλου επισκεπτών (4/5)

Για να προσθέσουμε στη ροή εξόρυξης το δεύτερο επίπεδο της θεματικής ιεραρχίας, τις υποκατηγορίες θεματικών όρων, θα πρέπει να προσθέσουμε ένα νέο χειριστή πηγών δεδομένων ο οποίος θα εισάγει στο μοντέλο τα δεδομένα του πίνακα SUBJECT_SUBCAT. Ο πίνακας SUBJECT_SUBCAT συσχετίζει τους θεματικούς όρους με τις υποκατηγορίες θεματικών όρων και θα αποτελέσει τον πίνακα ταξινομίας.

Πίνακας 6.15. Οι στήλες παιδιού και γονέα του πίνακα ταξινομίας SUBJECT_SUBCAT

Παιδί	Γονέας
S_ID	SUB_ID

Θα εισάγουμε επίσης έναν πίνακα αναζήτησης ονομάτων για την αντικατάσταση των κωδικών των υποκατηγοριών με τις αντίστοιχες ονομασίες τους. Ο πίνακας αυτός είναι ο πίνακας SUBCATEGORY. Για κάθε θεματική υποκατηγορία, φέρει τον κωδικό και την ονομασία της.

Εξετάζοντας τα δεδομένα του πίνακα SUBJECT_SUBCAT διαπιστώνουμε ότι ο περιορισμός που θέτει ο Intelligent Miner και ορίζει ότι μία κατηγορία δεν επιτρέπεται να είναι ούτε άμεσο ούτε έμμεσο μέλος του εαυτού της, παραβιάζεται. Για παράδειγμα, υπάρχει κωδικός S_ID με τιμή 26 και κωδικός SUB_ID με τιμή 26. Για να αντιμετωπίσουμε το πρόβλημα αυτό, επιλέγουμε τον αριθμό 1.000.000 τον οποίο και θα προσθέσουμε σε όλες τις τιμές SUB_ID του πίνακα SUBJECT_SUBCAT.

Στο σημείο αυτό εξηγείται ο λόγος για τον οποίο επιλέξαμε να προσθέσουμε στη στήλη S_ID του πίνακα ταξινομίας J_SUBJECT (και αντίστοιχα του πίνακα αναζήτησης ονομάτων SUBJECT) τον αριθμό 10.000.000. Δεδομένου ότι στον πίνακα ταξινομίας SUBJECT_SUBCAT συνυπάρχουν η προσαυξημένη κατά 10.000.000 στήλη S_ID και η στήλη SUB_ID, θα πρέπει να λάβουμε τα κατάλληλα μέτρα ώστε να μην υπάρχουν ίδιες τιμές στις δύο στήλες. Πρέπει λοιπόν να βρούμε έναν αριθμό αρκετά μακριά από τον αριθμό 10.000.000 τον οποίο θα προσθέσουμε σε όλες τις τιμές της στήλης SUB_ID. Γι'αυτό επιλέγουμε τον αριθμό 1.000.000. Καταφέρνουμε με αυτό τον τρόπο να διαφοροποιήσουμε τις τιμές των στηλών S_ID και SUB_ID με επιτυχία.

Η αλλαγή στις τιμές του πεδίου SUB_ID θα γίνει σε εικονικό πίνακα και όχι στη βάση δεδομένων. Για αυτό το λόγο, θα εισάγουμε στη ροή εξόρυξης ένα χειριστή <Select List> στον οποίο θα συνδέσουμε τον πίνακα SUBJECT_SUBCAT. Ο χειριστής αυτός θα παράγει στην έξοδό του έναν εικονικό πίνακα με δύο στήλες: S_ID και SUB_ID, με όλες τις τιμές του τελευταίου να είναι αυξημένες κατά 1.000.000.

Η ίδια ακριβώς αλλαγή θα πρέπει να γίνει και στη στήλη SUB_ID του πίνακα SUBCATEGORY ώστε να μη χαλάσει η αντιστοίχιση των ονομάτων με τους κωδικούς των θεματικών υποκατηγοριών. Θα χρησιμοποιήσουμε ένα δεύτερο χειριστή <Select List> ο οποίος θα τροποποιήσει ανάλογα τις τιμές του πεδίου SUB_ID του πίνακα SUBCATEGORY.

Τα βήματα για την προετοιμασία των παραπάνω ενεργειών είναι όμοια με αυτά που ακολουθήσαμε στην προσθήκη του επιπέδου θεματικών όρων. Γι'αυτό και δε θα τα περιγράψουμε αναλυτικά. Θα αναφέρουμε μόνο ορισμένα σημεία στα οποία υπάρχουν μικρές διαφορές.

1. Στον χειριστή <Select List> στον οποίο συνδέεται ο πίνακας ταξινομίας SUBJECT_SUBCAT, θα πρέπει να προστεθεί ο αριθμός 1.000.000 σε όλες τις τιμές της στήλης SUB_ID και ο αριθμός 10.000.000 σε όλες τις εγγραφές της στήλης S_ID.

2. Στον χειριστή <Select List> στον οποίο συνδέεται ο πίνακας αναζήτησης ονομάτων SUBCATEGORY, θα πρέπει να προστεθεί ο αριθμός 1.000.000 σε όλες τις τιμές της στήλης SUB_ID και το πρόθεμα 'Subc' πριν από τα ονόματα των υποκατηγοριών. Το πρόθεμα αυτό θα ξεχωρίζει τα αντικείμενα που αποτελούν θεματικές υποκατηγορίες.

3. Στον χειριστή [Journal Affinities] θα πρέπει να προσθέσουμε εκτός από μία νέα πόρτα εισόδου για τα δεδομένα αναζήτησης ονομάτων (Names3) και μία νέα πόρτα εισόδου για τα δεδομένα ταξινομίας (Category1).

3.3.2.1. Εκτέλεση της ροής εξόρυξης και προβολή του μοντέλου

Αφού εκτελέσουμε τη ροή εξόρυξης, το μοντέλο που παράγεται ανοίγει απευθείας στον IM Visualizer. Ένα δείγμα των κανόνων φαίνεται στο ακόλουθο σχήμα:

Rule	▼ Support	Confidence	Lift	Absolute Support
[Subc: Biology (General)] +[Subc: Physiology] ==> [Subc: Chemical technology]	0,7382%	66,3931%	8,83	38.022
[Subc: Physiology] +[Subc: Chemical technology] ==> [Subc: Biology (General)]	0,7382%	80,8772%	5,96	38.022
[Subc: Biology (General)] +[Subc: Chemical technology] ==> [Subc: Physiology]	0,7382%	88,5076%	6,68	38.022
[Subc: Biology (General)] +[Subc: Physiology] ==> [Subc: Medicine (General)]	0,7364%	66,2272%	7,21	37.927
[Subc: Physiology] +[Subc: Medicine (General)] ==> [Subc: Biology (General)]	0,7364%	85,8951%	6,33	37.927
[Subc: Biology (General)] +[Subc: Medicine (General)] ==> [Subc: Physiology]	0,7364%	73,5875%	5,55	37.927
[Subc: Biology (General)] +[Subc: Physiology] ==> [Subc: Internal medicine]	0,7278%	65,4519%	3,88	37.483
[Subc: Internal medicine] +[Subc: Biology (General)] ==> [Subc: Physiology]	0,7278%	78,6137%	5,93	37.483
[Subc: Internal medicine] +[Subc: Physiology] ==> [Subc: Biology (General)]	0,7278%	54,5326%	4,02	37.483
[Subc: Internal medicine] +[Subc: Therapeutics. Pharmacology] ==> [Subc: Phy...	0,7002%	72,1300%	5,44	36.065

Σχήμα 6.26. Δείγμα κανόνων συσχετίσεων μεταξύ θεματικών υποκατηγοριών

Στους κανόνες φαίνεται το πρόθεμα 'Subc' το οποίο ξεχωρίζει τα αντικείμενα που αποτελούν τις θεματικές υποκατηγορίες από όλα τα υπόλοιπα αντικείμενα (θεματικούς όρους και συγγράμματα). Και πάλι, οι τιμές support των κανόνων έχουν αυξηθεί σε σχέση με τους κανόνες του προηγούμενου μοντέλου.

Μερικοί κανόνες που περιλαμβάνουν θεματικούς όρους και θεματικές υποκατηγορίες φαίνονται στο ακόλουθο σχήμα:

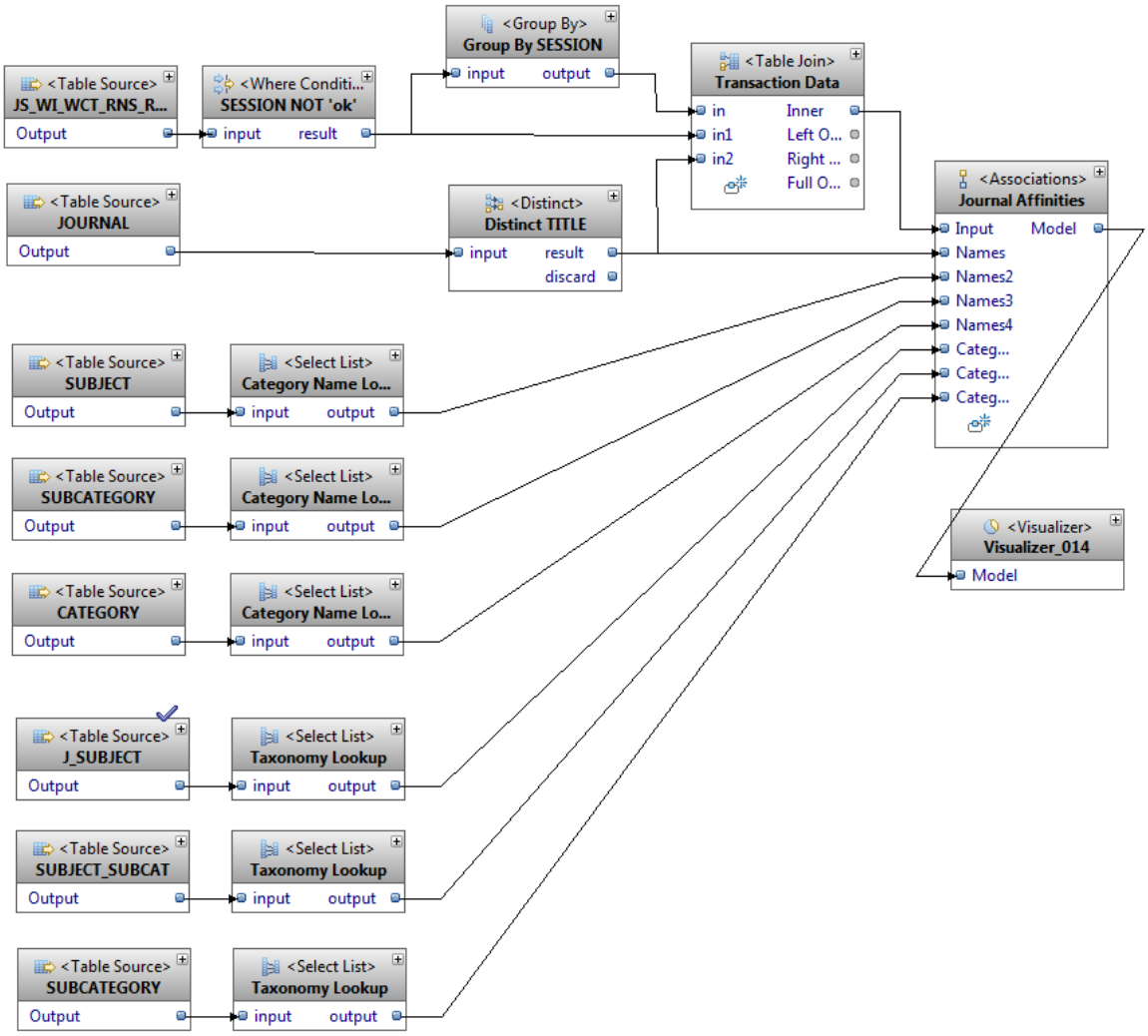
Rule	Support	Confidence	Lift	Absolute Support
[Subc: Physiology] +[Subj: Food -- Analysis] ==> [Subc: Public aspects of medicine]	0,0711%	84,5692%	10,39	3.661
[Subc: Public aspects of medicine] +[Subj: Food -- Analysis] ==> [Subc: Physiology]	0,0711%	96,6984%	7,29	3.661
[Subc: Microbiology] +[Subj: Food] ==> [Subc: Agriculture (General)]	0,0662%	96,5996%	45,17	3.409
[Subc: Agriculture (General)] +[Subj: Food] ==> [Subc: Microbiology]	0,0662%	93,5254%	21,27	3.409
[Subc: Internal medicine] +[Subj: Biochemistry] ==> [Subc: Medicine (General)]	0,0599%	76,0907%	8,29	3.087
[Subc: Medicine (General)] +[Subj: Biochemistry] ==> [Subc: Internal medicine]	0,0599%	71,6408%	4,25	3.087
[Subc: Agriculture (General)] +[Subj: Food -- Composition] ==> [Subc: Microbiology]	0,0592%	100,0000%	22,74	3.050
[Subc: Microbiology] +[Subj: Food -- Composition] ==> [Subc: Agriculture (General)]	0,0592%	100,0000%	46,76	3.050
[Subc: Chemical technology] +[Subj: Food -- Composition] ==> [Subc: Microbiology]	0,0592%	100,0000%	22,74	3.050
[Subc: Agriculture (General)] +[Subj: Food -- Composition] ==> [Subc: Chemical technology]	0,0592%	100,0000%	13,29	3.050

Σχήμα 6.27. Μερικοί κανόνες που συσχετίζουν θεματικούς όρους με θεματικές υποκατηγορίες

Στη συνέχεια θα προσθέσουμε στη ροή εξόρυξης και το τελευταίο επίπεδο της θεματικής ιεραρχίας, τις θεματικές κατηγορίες.

3.3.3. Προσθήκη επιπέδου θεματικών κατηγοριών

Με την ολοκλήρωση της προσθήκης του επιπέδου θεματικών υποκατηγοριών στις πληροφορίες ταξινομίας του μοντέλου, η ροή εξόρυξης θα μοιάζει με την ακόλουθη:



Σχήμα 6.28. Η ροή εξόρυξης για τη δημιουργία του μοντέλου επισκεπτών (5/5)

Για την προσθήκη στη ροή εξόρυξης του τρίτου επιπέδου της θεματική ιεραρχίας, θα χρειαστούμε τους ακόλουθους πίνακες:

Πίνακας ταξινομίας
SUBCATEGORY

Πίνακας αναζήτησης ονομάτων
CATEGORY

Πίνακας 6.16. Οι στήλες παιδιού και γονέα του πίνακα ταξινομίας SUBCATEGORY

Παιδί	Γονέας
SUB_ID	CAT_ID

Ο περιορισμός που θέτει ο Intelligent Miner για τα δεδομένα του πίνακα ταξινομίας παραβιάζεται και στην περίπτωση του πίνακα SUBCATEGORY. Υπάρχει κωδικός SUB_ID με τιμή 2 και κωδικός CAT_ID με τιμή 2. Για να αντιμετωπίσουμε το πρόβλημα, θα προσθέσουμε τον αριθμό 100.000 σε όλες τις τιμές της στήλης CAT_ID του πίνακα SUBCATEGORY. Επίσης, για να είναι συμβατοί οι κωδικοί SUB_ID του πίνακα SUBCATEGORY με τους αντίστοιχους κωδικούς του πίνακα SUBJECT_SUBCAT (οι οποίοι προσαυξήθηκαν στο προηγούμενο βήμα κατά 1.000.000), θα πρέπει να προσθέσουμε στους πρώτους τον αριθμό 1.000.000.

Ο λόγος για τον οποίο επιλέγουμε τον αριθμό 100.000 είναι γιατί τον χωρίζει μία ασφαλής απόσταση από τον αριθμό 1.000.000. Η επιλογή αυτή εξασφαλίζει ότι δε θα υπάρχουν ίδιες τιμές στα πεδία SUB_ID και CAT_ID. Συνολικά έως τώρα, στα βήματα για την προσθήκη πληροφοριών ταξινομίας στη ροή εξόρυξης, έχουμε κάνει τις ακόλουθες αλλαγές στις τιμές των πεδίων S_ID, SUB_ID και CAT_ID:

Πίνακας 6.17. Προσαυξήσεις τιμών των πεδίων S_ID, SUB_ID και CAT_ID

Πεδίο	Σταθερά προσαύξησης
S_ID	10.000.000
SUB_ID	1.000.000
CAT_ID	100.000

Με τις προσαυξήσεις αυτές είμαστε σίγουροι ότι δεν υπάρχουν κοινές τιμές στα τρία πεδία.

Για την υλοποίηση των παραπάνω στη ροή εξόρυξης θα χρειαστούμε:

- Δύο χειριστές πηγών δεδομένων <Table Source> για τους πίνακες SUBCATEGORY και CATEGORY.
- Δύο χειριστές επιλογής <Select List> οι οποίοι θα δημιουργήσουν τους εικονικούς πίνακες που θα τροφοδοτήσουν τον αλγόριθμο εξόρυξης.

Τα βήματα για την προετοιμασία των παραπάνω ενεργειών είναι όμοια με αυτά που ακολουθήσαμε στην προσθήκη του επιπέδου θεματικών υποκατηγοριών. Ορισμένα σημεία στα οποία υπάρχουν διαφορές είναι:

1. Στον χειριστή <Select List> στον οποίο συνδέεται ο πίνακας ταξινομίας SUBCATEGORY, θα πρέπει να προστεθεί ο αριθμός 100.000 σε όλες τις τιμές της στήλης CAT_ID και ο αριθμός 1.000.000 σε όλες τις εγγραφές της στήλης SUB_ID.

2. Στον χειριστή <Select List> στον οποίο συνδέεται ο πίνακας αναζήτησης ονομάτων CATEGORY, θα πρέπει να προστεθεί ο αριθμός 100.000 σε όλες τις τιμές της στήλης CAT_ID και το πρόθεμα 'Cat' πριν από τα ονόματα των κατηγοριών. Το πρόθεμα αυτό θα ξεχωρίζει τα αντικείμενα που αποτελούν θεματικές κατηγορίες.

3. Στον χειριστή [Journal Affinities] θα πρέπει να προσθέσουμε δύο νέες πόρτες εισόδου. Μία για τα δεδομένα αναζήτησης ονομάτων (Names4) και μία για τα δεδομένα ταξινομίας (Category2).

3.3.3.1. Εκτέλεση της ροής εξόρυξης και προβολή του μοντέλου

Αφού εκτελέσουμε τη ροή εξόρυξης, το μοντέλο που παράγεται ανοίγει απευθείας στον IM Visualizer. Ένα δείγμα των κανόνων φαίνεται στο ακόλουθο σχήμα:

Rule	▼ Support	Confidence	Lift	Absolute Support
[Cat: Science] +[Cat: Medicine] ==> [Cat: Technology]	1,3188%	50,6023%	2,00	67.924
[Cat: Science] +[Cat: Technology] ==> [Cat: Medicine]	1,3188%	64,5432%	2,06	67.924
[Cat: Medicine] +[Cat: Technology] ==> [Cat: Science]	1,3188%	95,8336%	2,35	67.924
[Cat: Medicine] +[Subc: Biology (General)] ==> [Subc: Physiology]	1,0018%	67,0095%	5,06	51.594
[Cat: Medicine] +[Subc: Physiology] ==> [Subc: Biology (General)]	1,0018%	56,7485%	4,19	51.594
[Subc: Biology (General)] +[Subc: Physiology] ==> [Cat: Medicine]	1,0018%	90,0922%	2,87	51.594
[Cat: Medicine] +[Subc: Biology (General)] ==> [Cat: Technology]	0,9957%	66,6030%	2,63	51.281
[Cat: Technology] +[Subc: Biology (General)] ==> [Cat: Medicine]	0,9957%	85,5396%	2,73	51.281
[Cat: Medicine] +[Cat: Technology] ==> [Subc: Biology (General)]	0,9957%	72,3521%	5,34	51.281
[Cat: Medicine] +[Cat: Technology] ==> [Subc: Physiology]	0,9132%	66,3558%	5,01	47.031

Σχήμα 6.29. Δείγμα κανόνων συσχετίσεων μεταξύ κατηγοριών και υποκατηγοριών

Στους κανόνες φαίνεται το πρόθεμα 'Cat' το οποίο ξεχωρίζει τα αντικείμενα που αποτελούν τις θεματικές κατηγορίες. Και πάλι, οι τιμές support των κανόνων έχουν αυξηθεί σε σχέση με τους κανόνες των προηγούμενων μοντέλων. Η αύξηση αυτή είναι λογική. Όσο περισσότερο γενικεύουμε τους κανόνες τόσο πιο πολύ μεγαλώνει η υποστήριξή τους.

Τα στατιστικά του μοντέλου φαίνονται στο ακόλουθο σχήμα:

▼ Global Statistics	
Number of transactions:	5.150.354
Average number of items per transactions:	1,12
Maximum number of items per transactions:	45
Number of item sets:	10.061
Number of singleton item sets:	80
Number of item sets used in rules:	1.094
Minimum rule support:	0,00%
Minimum rule confidence:	25,00%
Maximum rule length:	3
▼ Statistics for Visible Objects	
Visible rules:	9.994
Visible item sets:	10.061

Σχήμα 6.30. Στατιστικά του μοντέλου εξόρυξης των επισκεπτών

3.4. Αξιολόγηση του μοντέλου επισκεπτών

Το μοντέλο επισκεπτών πέρασε από διάφορες φάσεις μέχρι να φτάσει στην τελική του μορφή. Αρχικά, οι κανόνες του μοντέλου εξέφραζαν συσχετίσεις μόνο μεταξύ των συγγραμμάτων ενώ στη συνέχεια προσθέσαμε στη ροή εξόρυξης πληροφορίες ταξινομίας και γενικεύσαμε τους συσχετισμούς των αντικειμένων. Πλέον, ως αντικείμενα δε θεωρούνται μόνο τα συγγράμματα αλλά και οι κατηγορίες της θεματικής ιεραρχίας (οι θεματικοί όροι, οι θεματικές υποκατηγορίες και οι θεματικές κατηγορίες).

Η γενίκευση των κανόνων μάς δίνει τη δυνατότητα να μελετήσουμε τις συσχετίσεις που κρύβονται στα δεδομένα χρήσης του HEAL-Link με αποτελεσματικότερο τρόπο. Ένας κανόνας ο οποίος μάς πληροφορεί για τις προτιμήσεις των χρηστών σε επίπεδο θεματικών υποκατηγοριών μπορεί να είναι πολύ πιο χρήσιμος από έναν κανόνα ο οποίος εκφράζει τις προτιμήσεις των χρηστών σε συγκεκριμένους τίτλους συγγραμμάτων. Σε κάθε περίπτωση, η γενίκευση των κανόνων δε σημαίνει ότι δε μας ενδιαφέρουν οι πιο ειδικές περιπτώσεις. Εκείνο το οποίο θέλουμε είναι ένα μοντέλο ικανό να μας πληροφορεί για τις προτιμήσεις των χρηστών, όσο ειδικές ή γενικές και αν είναι. Χρησιμοποιώντας πληροφορίες ταξινομίας, ο στόχος αυτός επιτυγχάνεται σε μεγάλο βαθμό.

Ανάλογα με το πόσο επιθυμούμε να γενικεύσουμε ή να εμβαθύνουμε στις συσχετίσεις που κρύβονται στα δεδομένα, μπορούμε να χρησιμοποιήσουμε διαφορετικές εκδόσεις του μοντέλου. Από τις πιο ειδικές (σύγγραμμα ==> σύγγραμμα) έως τις πιο γενικές (κατηγορία ==> κατηγορία).

4. Δημιουργία μοντέλου εγγεγραμμένων μελών

Για τη δημιουργία του μοντέλου εγγεγραμμένων μελών, τα βήματα που θα ακολουθήσουμε είναι σχεδόν ίδια με αυτά της δημιουργίας του μοντέλου επισκεπτών. Οι διαφορές που θα συναντήσουμε είναι δύο:

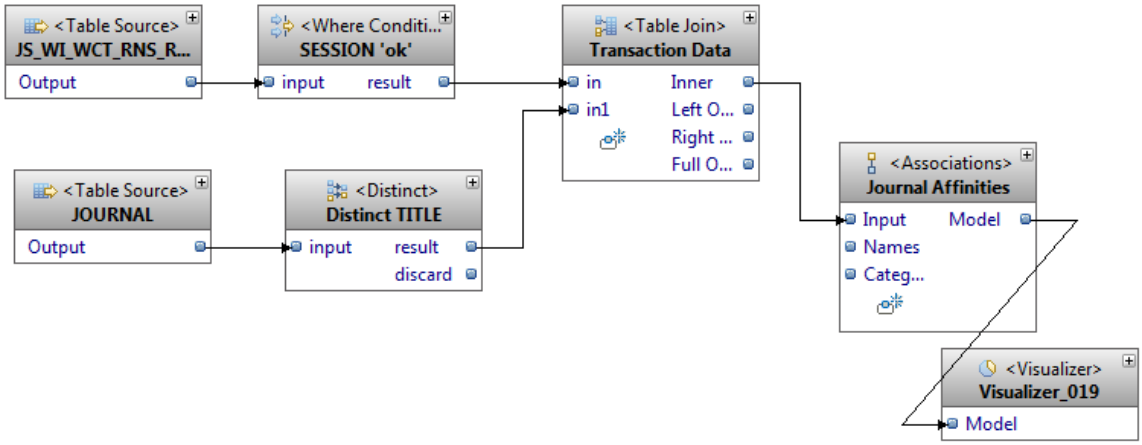
1. Το transaction id στο μοντέλο των εγγεγραμμένων μελών αποτελεί ο συνδυασμός των πεδίων EMAIL, YEAR, MONTH και DAY.

2. Δεν υπάρχουν δεδομένα συναλλαγών που να έχουν καταχωρηθεί στον πίνακα JS_WI_WCT_RNS_RCT_V2 από web-crawlers. Όλα τα δεδομένα των εγγεγραμμένων μελών προήλθαν από ανθρώπους και όχι από crawlers. Μπορούμε να είμαστε σίγουροι για αυτή τη διαπίστωση, αφού για να χρησιμοποιήσει το σύστημα ένας χρήστης-μέλος θα πρέπει πρώτα να κάνει login. Ένας crawler δε μπορεί να κάνει login σαν εγγεγραμμένο μέλος.

Και σε αυτή την περίπτωση, θα χρησιμοποιήσουμε πληροφορίες αναζήτησης ονομάτων και ταξινομίας για τη βελτίωση των παραγόμενων κανόνων.

Παρακάτω παρουσιάζονται οι πέντε φάσεις από τις οποίες περνάει η δημιουργία της ροής εξόρυξης και περιγράφονται τα κυριότερα σημεία τους.

4.1. Φάση πρώτη: Χωρίς πληροφορίες αναζήτησης ονομάτων και ταξινομίας



Σχήμα 6.31. Ροή εξόρυξης 1: Χωρίς πληροφορίες αναζήτησης ονομάτων και ταξινομίας

Στην πρώτη φάση η ροή εξόρυξης περιλαμβάνει μόνο τα δεδομένα συναλλαγών. Δεν συμπεριλαμβάνονται πληροφορίες αναζήτησης ονομάτων ούτε ταξινομίας. Τα αντικείμενα των παραγόμενων κανόνων αποτελούν μόνο ψηφιακά συγγράμματα και αναπαρίστανται με τους κωδικούς τους. Συνεπώς, είναι πολύ δύσκολο να διαβαστούν και να ερμηνευτούν.

Τα δεδομένα του πίνακα JS_WI_WCT_RNS_RCT_V2 εισάγονται στη ροή εξόρυξης μέσω ενός χειριστή <Table Source> και τροφοδοτούνται στον χειριστή [SESSION 'ok'] ο οποίος τα φιλτράρει και επιλέγει μόνο τις εγγραφές που έχουν στο πεδίο SESSION την τιμή 'ok'. Πρόκειται για τις εγγραφές του πίνακα JS_WI_WCT_RNS_RCT_V2 που καταχωρήθηκαν από εγγεγραμμένα μέλη. Η συνθήκη επιλογής των εγγραφών είναι:

```
SESSION LIKE 'ok'
```

Ο εικονικός πίνακας που δημιουργείται τροφοδοτείται στον χειριστή [Transaction Data] που είναι υπεύθυνος να συνθέσει το transaction id των συναλλαγών (EMAIL + YEAR + MONTH + DAY) και να αντικαταστήσει τους τίτλους των συγγραμμάτων με τους κωδικούς τους (J_ID). Οι κωδικοί των συγγραμμάτων προέρχονται από τον χειριστή [Distinct TITLE] ο οποίος αφαιρεί τις επαναλήψεις των ίδιων τίτλων του πίνακα JOURNAL. Αναφέραμε ήδη ότι μπορούμε άφοβα να αφαιρέσουμε τις επαναλήψεις των ίδιων τίτλων συγγραμμάτων από τον πίνακα

JOURNAL, καθώς όλες οι επαναλήψεις του ίδιου συγγράμματος ακολουθούν την ίδια διαδρομή στη θεματική ιεραρχία.

Η σύνθεση του transaction id γίνεται συνενώνοντας τα πεδία EMAIL, YEAR, MONTH και DAY με τη συνάρτηση CONCAT (||). Προηγουμένως, οι αριθμητικές τιμές των πεδίων (YEAR, MONTH και DAY) μετατρέπονται σε χαρακτήρες με τη συνάρτηση CHAR():

```
"IN_011_0_06"."EMAIL" || CHAR("IN_011_0_06"."YEAR") || CHAR("IN_011_0_06"."MONTH") ||
CHAR("IN_011_0_06"."DAY")
```

Στη στήλη Column Name του πίνακα Result Columns (ακριβώς δεξιά από τη στήλη Expression στην οποία ορίσαμε την παραπάνω έκφραση), καταχωρούμε την τιμή "EMAIL_Y_M_D". Αυτός θα είναι ο τίτλος της στήλης transaction id.

Η συνθήκη της σύζευξης που υλοποιεί ο χειριστής [Transaction Data] είναι:

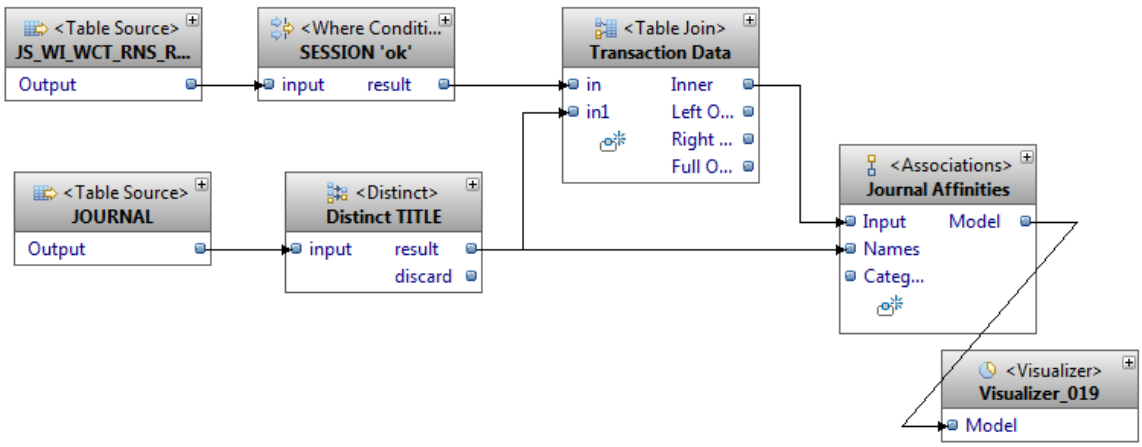
```
IN_011_0_06.JOURNAL = IN1_011_1_06.TITLE
```

Ο εικονικός πίνακας που παράγεται από τον χειριστή [Transaction Data] και αποτελείται από δύο στήλες (EMAIL +Y+M+D και J_ID) τροφοδοτείται στον χειριστή παραγωγής συσχετίσεων [Journal Affinities]. Οι ρυθμίσεις του τελευταίου φαίνονται στον ακόλουθο πίνακα:

Πίνακας 6.18. Οι ρυθμίσεις του χειριστή [Journal Affinities] στο μοντέλο εγγεγραμμένων μελών

Σελίδα	Ρυθμίσεις
Model Name	`MEMBER JOURNAL AFFINITIES`
Mining Settings	Journal Affinities Group column: EMAIL_Y_M_D Maximum rule length: 3 Maximum number of rules: 10000 Minimum confidence (%): 25 Minimum support (%): 0 Number of Bins: 5

4.2. Φάση δεύτερη: Πληροφορίες αναζήτησης ονομάτων, χωρίς ταξινομία



Σχήμα 6.32. Ροή εξόρυξης 2: Πληροφορίες αναζήτησης ονομάτων, χωρίς ταξινομία

Στη δεύτερη φάση προσθέτουμε στο μοντέλο πληροφορίες αναζήτησης ονομάτων για την παραγωγή ευανάγνωστων κανόνων. Ο εικονικός πίνακας που παράγει ο χειριστής [Distinct TITLE] περιέχει για κάθε κωδικό συγγράμματος την αντίστοιχη ονομασία του. Συνδέουμε την έξοδο **result** του χειριστή [Distinct TITLE] στην είσοδο **Names** του χειριστή [Journal Affinities]. Στη συνέχεια, ρυθμίζουμε το χειριστή [Journal Affinities] ως εξής:

- Στη σελίδα Name Maps, ορίζουμε ποιες στήλες του πίνακα αναζήτησης ονομάτων θα χρησιμοποιήσει ο χειριστής για να αντικαταστήσει τους κωδικούς με τους τίτλους των συγγραμμάτων. Οι επιλογές φαίνονται στον ακόλουθο πίνακα:

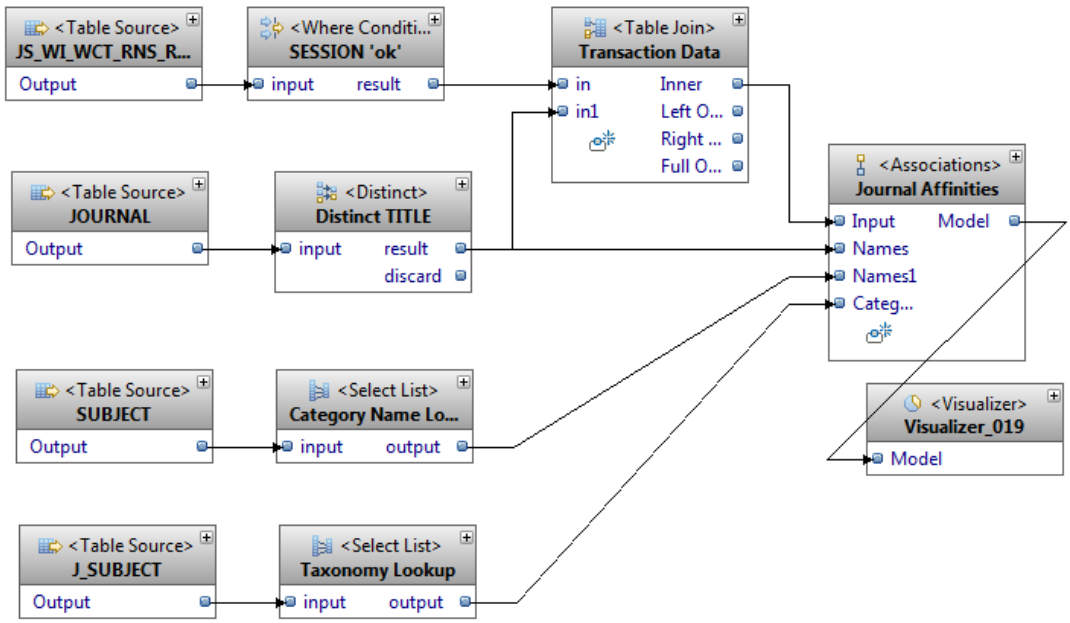
Πίνακας 6.19. Ρυθμίσεις αντιστοίχισης ονομάτων στη ροή εξόρυξης εγγεγραμμένων μελών

Map Name	Item Id Column	Item Name Column
Names	J_ID	TITLE

  Με τις παραπάνω ρυθμίσεις θα γνωρίζει ο Intelligent Miner ότι πρέπει να αντικαταστήσει τις τιμές της στήλης J_ID με τις αντίστοιχες τιμές της στήλης TITLE, για κάθε εγγραφή που δέχεται στην πόρτα εισόδου (Input).

- Στη σελίδα Column Properties, επιλέγουμε την τιμή **Names** στη στήλη **Name Mapping**.

4.3. Φάση τρίτη: Πληροφορίες αναζήτησης ονομάτων και ταξινομία πρώτου επιπέδου



Σχήμα 6.33. Ροή εξόρυξης 3: Πληροφορίες αναζήτησης ονομάτων και ταξινομία πρώτου επιπέδου

Στην τρίτη φάση προσθέτουμε δύο χειριστές πηγών δεδομένων οι οποίοι εισάγουν στο μοντέλο τα δεδομένα των πινάκων J_SUBJECT και SUBJECT. Ο πίνακας J_SUBJECT θα χρησιμοποιηθεί ως πίνακας ταξινομίας και ο πίνακας SUBJECT ως πίνακας αναζήτησης ονομάτων για τους θεματικούς όρους του J_SUBJECT. Οι δύο αυτοί πίνακες εισάγουν στο μοντέλο πληροφορίες ταξινομίας πρώτου επιπέδου. Αυτό σημαίνει ότι οι παραγόμενοι κανόνες θα μπορούν να περιλαμβάνουν ως αντικείμενα και τους θεματικούς όρους.

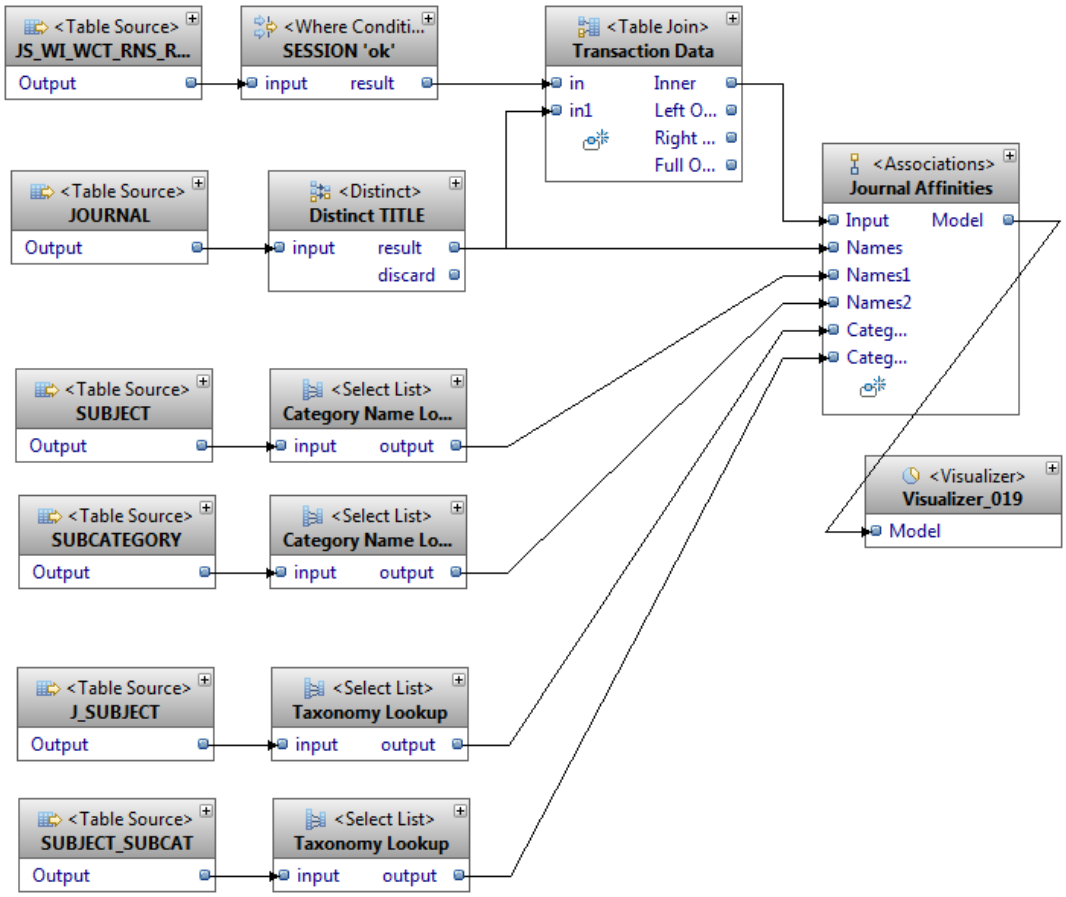
Επίσης, εισάγουμε δύο χειριστές <Select List> οι οποίοι θα δημιουργήσουν δύο εικονικούς πίνακες από τα δεδομένα των J_SUBJECT και SUBJECT. Αφού προεπεξεργαστούν τα δεδομένα, θα τα τροφοδοτήσουν στις πόρτες εισόδου **Names1** και **Category** του χειριστή [Journal Affinities].

Η προεπεξεργασία περιλαμβάνει την πρόσθεση του αριθμού 10.000.000 σε όλες τις τιμές της στήλης J_SUBJECT.S_ID. Η πρόσθεση αυτή γίνεται για να ικανοποιηθεί ο περιορισμός που θέτει ο Intelligent Miner

και ορίζει ότι μία κατηγορία δεν επιτρέπεται να είναι ούτε άμεσο ούτε έμμεσο μέλος του εαυτού της. Δηλαδή, δεν πρέπει να υπάρχουν στον πίνακα ταξινομίας ίδιες τιμές στις στήλες J_ID και S_ID οι οποίες αποτελούν στην ιεραρχία κατηγοριών παιδί και γονέα αντίστοιχα. Ο ίδιος αριθμός θα πρέπει να προστεθεί και σε όλες τις τιμές της στήλης SUBJECT.S_ID έτσι ώστε να διατηρηθεί η αντιστοίχηση ονομάτων των θεματικών όρων.

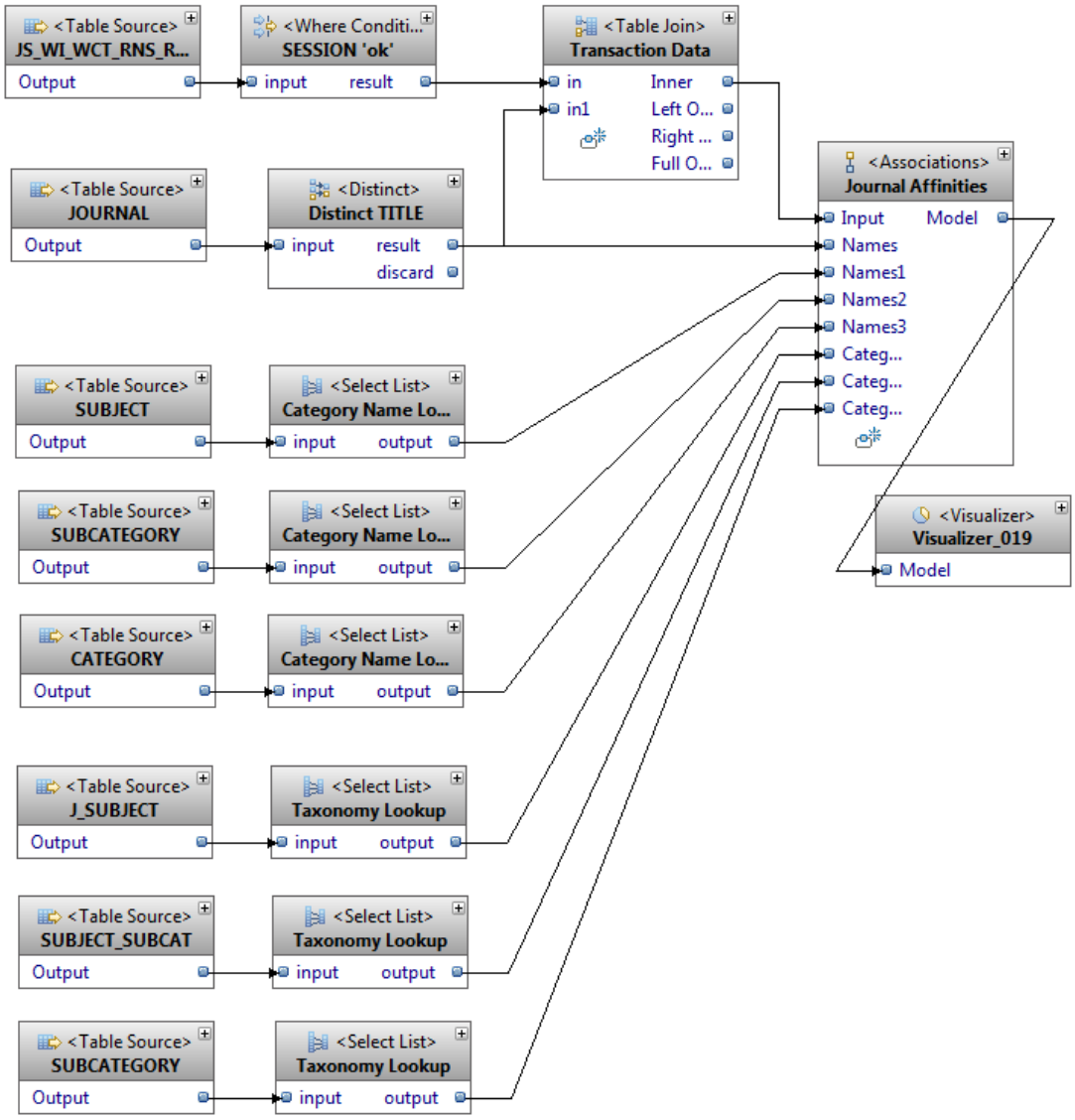
Τέλος, θα πρέπει να ρυθμίσουμε κατάλληλα τον χειριστή [Journal Affinities] ώστε να γνωρίζει πως να χρησιμοποιήσει τις πληροφορίες ταξινομίας. Οι οδηγίες ρύθμισης είναι ακριβώς ίδιες με αυτές που ακολουθήσαμε στη δημιουργία του μοντέλου επισκεπτών.

4.4. Φάση τέταρτη και πέμπτη: Πληροφορίες αναζήτησης ονομάτων και ταξινομία πρώτου, δεύτερου και τρίτου επιπέδου



Σχήμα 6.34. Ροή εξόρυξης 4: Πληροφορίες αναζήτησης ονομάτων και ταξινομία πρώτου και δεύτερου επιπέδου

Στην τέταρτη και πέμπτη φάση εισάγουμε στη ροή εξόρυξης τους πίνακες SUBJECT_SUBCAT, SUBCATEGORY και CATEGORY προσθέτοντας στις πληροφορίες ταξινομίας το δεύτερο και τρίτο επίπεδο της θεματικής ιεραρχίας των συγγραμμάτων, δηλαδή τις θεματικές υποκατηγορίες και κατηγορίες. Οι οδηγίες υλοποίησης των παραπάνω είναι ακριβώς ίδιες με αυτές που ακολουθήσαμε στη δημιουργία του μοντέλου επισκεπτών.



Σχήμα 6.35. Ροή εξόρυξης 5: Πληροφορίες αναζήτησης ονομάτων και ταξινομία πρώτου, δεύτερου και τρίτου επιπέδου

4.5. Αξιολόγηση του μοντέλου εγγεγραμμένων μελών

Το μοντέλο εγγεγραμμένων μελών σίγουρα παράγει κανόνες με πολύ καλύτερη υποστήριξη. Αυτό οφείλεται στο γεγονός ότι το πλήθος των συναλλαγών είναι σχετικά μικρό (16.197). Σε ότι αφορά τη γνώση που μας παρέχει, θα μπορούσαμε να πούμε ότι παράγει κανόνες παρόμοιους με αυτούς του μοντέλου επισκεπτών. Συγκρίνοντας τα αποτελέσματα των δύο μοντέλων σε όλα τα επίπεδα ταξινομίας που χρησιμοποιήθηκαν, διαπιστώνουμε ότι δεν υπάρχουν ιδιαίτερες διαφορές. Βασική διαφορά των δύο είναι η βάση των συναλλαγών πάνω στην οποία παρήγαγαν τους κανόνες. Το μοντέλο επισκεπτών σίγουρα αναλύει πολύ περισσότερες συναλλαγές.

Παρακάτω παρουσιάζονται δείγματα των κανόνων που παρήγαγε το μοντέλο εγγεγραμμένων μελών σε όλα τα επίπεδα ταξινομίας. Επίσης, απεικονίζονται τα στατιστικά του μοντέλου έτσι όπως διαμορφώθηκαν ύστερα από την προσθήκη και του τρίτου επιπέδου της θεματικής ιεραρχίας στις πληροφορίες ταξινομίας.

Rule	▼ Support	Confidence	Lift	Absolute Support
[British Journal of Haematology] ==> [Current Opinion in Hematology]	0,2964%	26,9663%	49,63	48
[Current Opinion in Hematology] ==> [British Journal of Haematology]	0,2964%	54,5455%	49,63	48
[Annals of Hematology] ==> [British Journal of Haematology]	0,2531%	48,8095%	44,41	41
[Forest Ecology and Management] ==> [Forestry]	0,1914%	47,6923%	175,56	31
[Forestry] ==> [Forest Ecology and Management]	0,1914%	70,4545%	175,56	31
[Environmental Education Research] ==> [Journal of Environmental Education, ...	0,1790%	40,8451%	98,74	29
[Journal of Environmental Education, The] ==> [Environmental Education Rese...	0,1790%	43,2836%	98,74	29
[Biochimica et Biophysica Acta (BBA) - Bioenergetics (incorporating Biochimica et ...	0,1605%	26,8041%	68,91	26
[Biochemistry] ==> [Biochimica et Biophysica Acta (BBA) - Bioenergetics (incorp...	0,1605%	41,2698%	68,91	26
[Biochimica et Biophysica Acta (BBA) - Bioenergetics (incorporating Biochimica et ...	0,1605%	26,8041%	33,65	26

Σχήμα 6.36. Δείγμα κανόνων μοντέλου εγγεγραμμένων μελών χωρίς ταξινομία

Rule	▼ Support	Confidence	Lift	Absolute Support
[Subj: Blood -- Diseases] ==> [Subj: Hematology]	0,7656%	83,7838%	41,00	124
[Subj: Hematology] ==> [Subj: Blood -- Diseases]	0,7656%	37,4622%	41,00	124
[Subj: Heart -- Diseases] ==> [Subj: Cardiology]	0,6297%	74,4526%	42,16	102
[Subj: Gynecology] ==> [Subj: Obstetrics]	0,6297%	68,4564%	65,22	102
[Subj: Obstetrics] ==> [Subj: Gynecology]	0,6297%	60,0000%	65,22	102
[Subj: Cardiology] ==> [Subj: Heart -- Diseases]	0,6297%	35,6643%	42,16	102
[Subj: Oncology] ==> [Subj: Tumors]	0,5804%	55,2941%	60,11	94
[Subj: Tumors] ==> [Subj: Oncology]	0,5804%	63,0872%	60,11	94
[Subj: Spine -- Abnormalities] ==> [Subj: Spine -- Diseases]	0,5557%	79,6460%	94,16	90
[Subj: Spine -- Diseases] ==> [Subj: Spine -- Abnormalities]	0,5557%	65,6934%	94,16	90

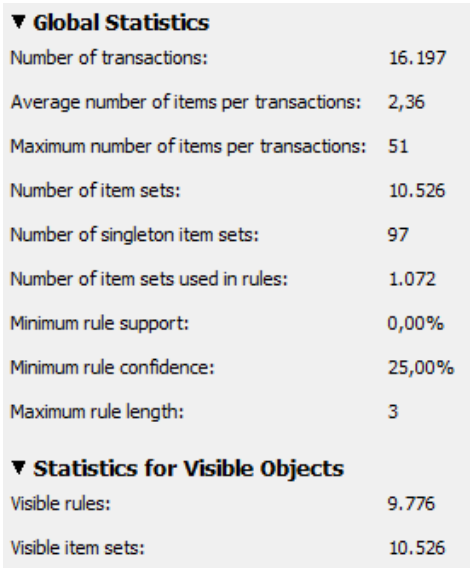
Σχήμα 6.37. Δείγμα κανόνων μοντέλου εγγεγραμμένων μελών με ταξινομία 1ου επιπέδου (θεματικοί όροι)

Rule	▼ Support	Confidence	Lift	Absolute Support
[Subc: Physiology] ==> [Subc: Internal medicine]	11,6503%	55,3860%	1,69	1.887
[Subc: Internal medicine] ==> [Subc: Physiology]	11,6503%	35,5702%	1,69	1.887
[Subc: Therapeutics. Pharmacology] ==> [Subc: Internal medicine]	8,8411%	60,6266%	1,85	1.432
[Subc: Internal medicine] ==> [Subc: Therapeutics. Pharmacology]	8,8411%	26,9934%	1,85	1.432
[Subc: Physiology] ==> [Subc: Biology (General)]	8,0138%	38,0980%	1,93	1.298
[Subc: Biology (General)] ==> [Subc: Physiology]	8,0138%	40,5498%	1,93	1.298
[Subc: Surgery] ==> [Subc: Internal medicine]	7,8224%	53,1014%	1,62	1.267
[Subc: Medicine (General)] ==> [Subc: Biology (General)]	7,7175%	51,9751%	2,63	1.250
[Subc: Biology (General)] ==> [Subc: Medicine (General)]	7,7175%	39,0503%	2,63	1.250
[Subc: Biology (General)] ==> [Subc: Internal medicine]	7,5631%	38,2693%	1,17	1.225

Σχήμα 6.38. Δείγμα κανόνων μοντέλου εγγεγραμμένων μελών με ταξινομία 1ου και 2ου επιπέδου (θεματικοί όροι, θεματικές υποκατηγορίες)

Rule	▼ Support	Confidence	Lift	Absolute Support
[Cat: Science] ==> [Cat: Medicine]	21,8312%	43,2962%	0,84	3.536
[Cat: Medicine] ==> [Cat: Science]	21,8312%	42,4490%	0,84	3.536
[Cat: Technology] ==> [Cat: Science]	16,9105%	53,7797%	1,07	2.739
[Cat: Science] ==> [Cat: Technology]	16,9105%	33,5374%	1,07	2.739
[Subc: Internal medicine] ==> [Cat: Science]	15,0522%	45,9566%	0,91	2.438
[Cat: Science] ==> [Subc: Internal medicine]	15,0522%	29,8518%	0,91	2.438
[Cat: Medicine] ==> [Subc: Physiology]	14,1508%	27,5150%	1,31	2.292
[Subc: Physiology] ==> [Cat: Medicine]	14,1508%	67,2733%	1,31	2.292
[Subc: Biology (General)] ==> [Cat: Medicine]	11,7367%	59,3877%	1,15	1.901
[Subc: Internal medicine] ==> [Subc: Physiology]	11,6503%	35,5702%	1,69	1.887

Σχήμα 6.39. Δείγμα κανόνων μοντέλου εγγεγραμμένων μελών με ταξινομία 1ου, 2ου και 3ου επιπέδου (θεματικοί όροι, θεματικές υποκατηγορίες, θεματικές κατηγορίες)



▼ **Global Statistics**

Number of transactions:	16.197
Average number of items per transactions:	2,36
Maximum number of items per transactions:	51
Number of item sets:	10.526
Number of singleton item sets:	97
Number of item sets used in rules:	1.072
Minimum rule support:	0,00%
Minimum rule confidence:	25,00%
Maximum rule length:	3

▼ **Statistics for Visible Objects**

Visible rules:	9.776
Visible item sets:	10.526

Σχήμα 6.40. Στατιστικά μοντέλου εγγεγραμμένων μελών με πληροφορίες ταξινομίας 1ου, 2ου και 3ου επιπέδου

Σε κάθε περίπτωση, εάν έπρεπε να επιλέξουμε ανάμεσα στα δύο μοντέλα για τη δημιουργία ενός συστήματος παραγωγής συστάσεων, η καλύτερη επιλογή θα ήταν το μοντέλο επισκεπτών κι αυτό γιατί η βάση συναλλαγών την οποία ανέλυσε είναι πολλές χιλιάδες φορές μεγαλύτερη από τη βάση συναλλαγών που ανέλυσε το μοντέλο εγγεγραμμένων μελών. Αυτό συνεπάγεται πιο έγκυρα και ακριβή αποτελέσματα τα οποία προσεγγίζουν περισσότερο την πραγματικότητα.

Στο επόμενο κεφάλαιο θα αναπτύξουμε μία απλή εφαρμογή στο περιβάλλον του Infosphere Warehouse Server η οποία θα διαχειρίζεται και θα εκτελεί τα παραπάνω μοντέλα ανά τακτά χρονικά διαστήματα, εξάγοντας παράλληλα τους παραγόμενους κανόνες σε πίνακες της βάσης δεδομένων και καθιστώντας τους με αυτό τον τρόπο προσβάσιμους σε τρίτες εφαρμογές.

Μία τέτοια εφαρμογή θα μπορούσε να είναι ένα σύστημα παραγωγής συστάσεων προς τους χρήστες της διαδικτυακής πύλης HEAL-Link. Με βάση τις στοχοποιήσεις-επιλογές συγγραμμάτων που κάνουν οι χρήστες, το σύστημα θα τους προτείνει συγγράμματα ή αντικείμενα της θεματική ιεραρχίας που είναι πιθανό να τους ενδιαφέρουν. Οι προτάσεις αυτές θα γίνονται με βάση τις γενικές προτιμήσεις όλων των χρηστών.

Κεφάλαιο 7. Ανάπτυξη εφαρμογής εξόρυξης

1. Εισαγωγή

Στο κεφάλαιο αυτό θα αναπτύξουμε μία εφαρμογή εξόρυξης (mining application) η οποία θα χρησιμοποιεί τη ροή εξόρυξης επισκεπτών που δημιουργήσαμε στη φάση μοντελοποίησης για να διατηρεί ενημερωμένους δύο πίνακες εντός των οποίων θα βρίσκονται αποθηκευμένοι οι κανόνες του μοντέλου.

Στο προηγούμενο κεφάλαιο, σχεδιάσαμε δύο ροές εξόρυξης οι οποίες δημιουργούν μοντέλα συσχετίσεων για την ανακάλυψη κρυμμένης γνώσης στα δεδομένα συναλλαγών του HEAL-Link portal, τα οποία καταχωρήθηκαν από επισκέπτες και εγγεγραμμένα μέλη. Επίσης, μελετήσαμε τα μοντέλα με τον IM Visualizer και επιχειρήσαμε να τα βελτιώσουμε. Τώρα, θα τοποθετήσουμε τους κανόνες συσχετίσεων που παρήγαγε το μοντέλο επισκεπτών σε πίνακες της βάσης δεδομένων ώστε να καταστούν προσβάσιμοι από τρίτες εφαρμογές.

Η διαδικτυακή εφαρμογή του HEAL-Link θα μπορούσε να χρησιμοποιήσει τους εξαχθέντες κανόνες για να προτείνει συγγράμματα, θεματικούς όρους, θεματικές υποκατηγορίες ή θεματικές κατηγορίες σε έναν χρήστη, ανάλογα με τις επιλογές που κάνει. Επίσης, η ομάδα διαχείρισης του διαδικτυακού τόπου θα μπορούσε να χρησιμοποιήσει μία εφαρμογή αναφορών (reporting application) η οποία θα διαβάζει τον πίνακα των κανόνων και θα αναφέρει στους διαχειριστές τις πιο αξιοσημείωτες συσχετίσεις, βοηθώντας τους με αυτόν τον τρόπο να χειριστούν καλύτερα τα περιεχόμενα της βάσης δεδομένων αλλά και να βελτιώσουν τη συμπεριφορά της διαδικτυακής πύλης απέναντι στους χρήστες.

Πρώτα θα χρησιμοποιήσουμε το Design Studio για να προσθέσουμε στη ροή εξόρυξης επισκεπτών περαιτέρω βήματα τα οποία θα εξάγουν τους κανόνες σε πίνακες της βάσης δεδομένων. Στη συνέχεια, θα δημιουργήσουμε μία εφαρμογή εξόρυξης (data warehouse application) κατασκευάζοντας μία ροή ελέγχου (control flow) που θα εκτελεί τη ροή εξόρυξης και τελικά θα πακετάρουμε τις δύο ροές σε ένα συμπιεσμένο αρχείο.

Μόλις ολοκληρωθεί η δημιουργία της εφαρμογής, θα χρησιμοποιήσουμε την κονσόλα διαχείρισης της πλατφόρμας Infosphere™ Warehouse για να εγκαταστήσουμε την εφαρμογή (το συμπιεσμένο αρχείο) στον Infosphere Warehouse Server και θα την προγραμματίσουμε ώστε να εκτελείται ανά τακτά χρονικά διαστήματα.

Με την ολοκλήρωση του κεφαλαίου, θα έχουμε στη διάθεσή μας τακτικά ενημερωνόμενους πίνακες κανόνων συσχετίσεων οι οποίοι θα είναι διαθέσιμοι σε οποιαδήποτε εφαρμογή ζητήσει πρόσβαση στο περιεχόμενό τους.

Τα βήματα που θα ακολουθήσουμε για την υλοποίηση των παραπάνω είναι:

1. Προσθήκη βημάτων στη ροή εξόρυξης για την εξαγωγή των κανόνων συσχετίσεων
2. Σχεδιασμός μίας απλής ροής ελέγχου (control flow) η οποία θα είναι υπεύθυνη να εκτελεί τη ροή εξόρυξης
3. Προετοιμασία μίας εφαρμογής εξόρυξης η οποία θα εγκατασταθεί στον Infosphere Warehouse Server και θα παράγει τους κανόνες συσχετίσεων
4. Εγκατάσταση της εφαρμογής εξόρυξης στον Infosphere Warehouse Server με τη βοήθεια της κονσόλας διαχείρισης της πλατφόρμας

2. Εξαγωγή του μοντέλου εξόρυξης

Στο τελευταίο βήμα της ροής εξόρυξης, οι κανόνες που παράγει ο χειριστής <Associations> κατευθύνονται στον χειριστή <Visualizer> ο οποίος εκκινεί τον IM Visualizer και μας επιτρέπει να μελετήσουμε το μοντέλο. Στο τρέχον βήμα, θα προσθέσουμε στη ροή εξόρυξης ενέργειες οι οποίες θα εκτελεστούν παράλληλα με τον χειριστή <Visualizer> και θα οργανώσουν τους κανόνες συσχετίσεων σε επιμέρους πεδία τα οποία θα αποθηκευτούν σε πίνακες της βάσης δεδομένων.

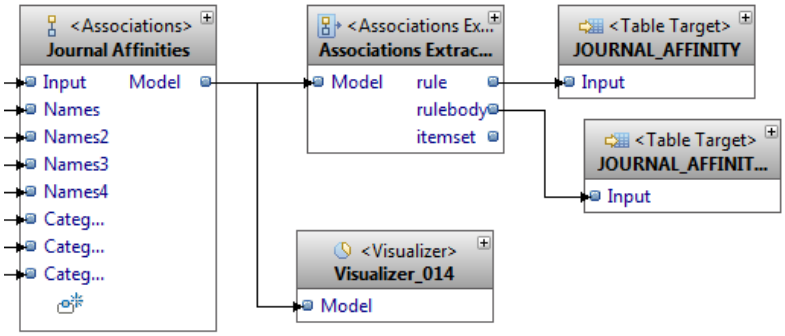
Συγκεκριμένα, θα εκτελέσουμε τις ακόλουθες ενέργειες:

1. Προσθήκη ενός χειριστή εξαγωγής κανόνων συσχετίσεων <Associations Extractor>

2. Δημιουργία ενός πίνακα προορισμού για τους κανόνες συσχετίσεων προσθέτοντας ένα χειριστή <Table Target>

3. Δημιουργία ενός πίνακα προορισμού για τα σώματα των κανόνων συσχετίσεων προσθέτοντας ένα χειριστή <Table Target>

Με την ολοκλήρωση των παραπάνω ενεργειών, η απόληξη της ροής εξόρυξης θα μοιάζει με αυτήν του ακόλουθου σχήματος:

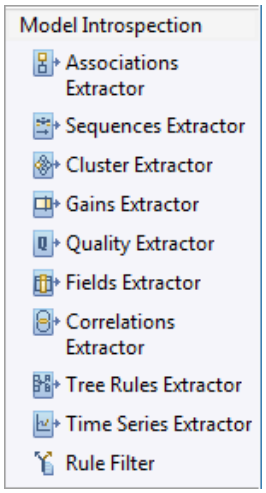


Σχήμα 7.1. Εξαγωγή των κανόνων συσχετίσεων σε πίνακες της βάσης δεδομένων

2.1. Προσθήκη χειριστή εξαγωγής κανόνων συσχετίσεων

Για να προσθέσουμε στη ροή εξόρυξης ένα χειριστή εξαγωγής κανόνων συσχετίσεων:

1. Από την παλέτα χειριστών στην δεξιά πλευρά του καμβά, επιλέγουμε τον χειριστή **Associations Extractor** (στην ομάδα Model Introspection) και τον τοποθετούμε στον καμβά.



Σχήμα 7.2. Ο χειριστής Associations Exctractor στην παλέτα χειριστών

2. Ορίζουμε τον χειριστή [Journal Affinities] ως είσοδο στον χειριστή <Associations Extractor>:

   - Κάνουμε κλικ στην πόρτα εξόδου **Model** του χειριστή [Journal Affinities] και στη συνέχεια στην πόρτα εισόδου **Model** του χειριστή <Associations Extractor>.

Κάθε πόρτα εξόδου του χειριστή <Associations Extractor> αντιπροσωπεύει ένα διαφορετικό πίνακα πληροφοριών του μοντέλου κανόνων συσχετίσεων. Η πόρτα εξόδου **rule** αντιπροσωπεύει έναν πίνακα ο οποίος περιλαμβάνει τα περισσότερα στοιχεία των κανόνων συσχετίσεων καθώς και πληροφορίες μέτρησης της ποιότητας των κανόνων. Ακολουθεί ένα παράδειγμα των εγγραφών του πίνακα στην έξοδο rule:

Πίνακας 7.1. Αντιπροσωπευτικό δείγμα των περιεχομένων του πίνακα που παράγεται στην έξοδο rule του χειριστή <Associations Extractor>

id	bodyid	head	headname	length	bodytext	support	confidence	lift
9.994	2	100024	Cat: Technology	3	[Cat: Science] [Cat: Medicine]	0,013188	0,5060232	1,9993806

Η εγγραφή περιλαμβάνει τις τιμές support, confidence και lift του κανόνα "χρήστες οι οποίοι στοχοποιούν συγγράμματα των κατηγοριών Science και Medicine τείνουν να στοχοποιούν και συγγράμματα της κατηγορίας Technology".

Ο πίνακας αυτός δεν περιλαμβάνει τους κωδικούς (ID) των αντικειμένων του σώματος του κανόνα. Ο λόγος για τον οποίο δεν είναι δυνατό να υπάρχει στον παραπάνω πίνακα στήλη ID για του κωδικούς των αντικειμένων του σώματος, είναι γιατί κάθε κανόνας μπορεί να περιέχει στο σώμα του περισσότερα από ένα αντικείμενα. Για την εφαρμογή παραγωγής συστάσεων, όμως, θέλουμε να έχουμε τη δυνατότητα να συμβουλευόμαστε κανόνες οι οποίοι σχετίζονται με τα συγγράμματα που στοχοποιούν οι χρήστες. Ο μόνος τρόπος για να εντοπίσουμε αυτούς τους κανόνες είναι να γνωρίζουμε τους κωδικούς των αντικειμένων στο σώμα τους ώστε να μπορέσουμε να τους συγκρίνουμε με τα αντικείμενα που έχει ήδη στοχοποιήσει ο χρήστης.

Για την αντιστοίχιση των συγγραμμάτων που έχει ήδη στοχοποιήσει ένας χρήστης, με τους κανόνες συσχετίσεων που θα μπορούσαν να παρέχουν συστάσεις, μπορούμε να χρησιμοποιήσουμε την πόρτα εξόδου rulebody του χειριστή <Associations Extractor>. Στον ακόλουθο πίνακα παρουσιάζεται ένα δείγμα των εγγραφών που παράγει η έξοδος rulebody.

Πίνακας 7.2. Αντιπροσωπευτικό δείγμα των εγγραφών του πίνακα που παράγεται στην έξοδο rulebody του χειριστή <Associations Extractor>

BODYID	ITEMNAME	ITEM
2	[Cat: Science]	100022
2	[Cat: Medicine]	100017

Οι δύο εγγραφές μαζί αποτελούν το σώμα του κανόνα.

Στη συνέχεια θα δημιουργήσουμε δύο πίνακες προορισμού στους οποίους θα αποθηκευτούν οι εγγραφές που παράγουν οι πόρτες εξόδου rule και rulebody του χειριστή <Associations Extractor>.

2.2. Δημιουργία πίνακα προορισμού για τους κανόνες συσχετίσεων

Για να δημιουργήσουμε τον πίνακα προορισμού JOURNAL_AFFINITY στον οποίο θα αποθηκευτούν οι κανόνες συσχετίσεων του μοντέλου, ακολουθούμε τα εξής βήματα:

- Κάνουμε δεξί κλικ στην πόρτα εξόδου **rule** του χειριστή <Associations Extractor> και επιλέγουμε **Create Suitable Table**.
- Στο αναδυόμενο παράθυρο, ορίζουμε τις ακόλουθες παραμέτρους:

Πίνακας 7.3. Παράμετροι δημιουργίας του πίνακα προορισμού JOURNAL_AFFINITY

Παράμετρος	Τιμή
Database / data model	Αποδοχή της προεπιλεγμένης τιμής.
Table name	JOURNAL_AFFINITY
Table schema	DB2ADMIN
Table space	Αποδοχή της προεπιλεγμένης τιμής.

Παράμετρος	Τιμή
Automatically create and connect to target operator	Ενεργοποίηση αυτής της επιλογής.

- Πατάμε το κουμπί Finish.

Το Design Studio δημιουργεί τον πίνακα JOURNAL_AFFINITY στο σχήμα DB2ADMIN και τοποθετεί στη ροή εξόρυξης τον κατάλληλο χειριστή <Table Target>.

Για να είμαστε σίγουροι ότι ο πίνακας JOURNAL_AFFINITY θα περιέχει πάντα τους τελευταίους και πιο ενημερωμένους κανόνες του μοντέλου, θα πρέπει να πούμε στον χειριστή [JOURNAL_AFFINITY] να διαγράφει τα περιεχόμενα του πίνακα προτού εισάγει νέα:

- Κάνοντας διπλό κλικ στον χειριστή <Table Target> του πίνακα JOURNAL_AFFINITY εμφανίζεται ο οδηγός ρύθμισής του. Στη σελίδα General, ενεργοποιούμε την επιλογή **Delete previous content**.

2.3. Δημιουργία πίνακα προορισμού για τα σώματα των κανόνων συσχετίσεων

Για να δημιουργήσουμε τον πίνακα προορισμού JOURNAL_AFFINITY_BODY στον οποίο θα αποθηκευτούν τα σώματα των κανόνων συσχετίσεων, ακολουθούμε τα εξής βήματα:

- Κάνουμε δεξί κλικ στην πόρτα εξόδου **rulebody** του χειριστή <Associations Extractor> και επιλέγουμε **Create Suitable Table**.

- Στο αναδυόμενο παράθυρο, ορίζουμε τις ακόλουθες παραμέτρους:

Πίνακας 7.4. Παράμετροι δημιουργίας του πίνακα προορισμού JOURNAL_AFFINITY_BODY

Παράμετρος	Τιμή
Database / data model	Αποδοχή της προεπιλεγμένης τιμής.
Table name	JOURNAL_AFFINITY_BODY
Table schema	DB2ADMIN
Table space	Αποδοχή της προεπιλεγμένης τιμής.
Automatically create and connect to target operator	Ενεργοποίηση αυτής της επιλογής.

- Πατάμε το κουμπί Finish.

Το Design Studio δημιουργεί τον πίνακα JOURNAL_AFFINITY_BODY στο σχήμα DB2ADMIN και τοποθετεί στη ροή εξόρυξης τον κατάλληλο χειριστή <Table Target>.

- Κάνοντας διπλό κλικ στον χειριστή <Table Target> του πίνακα JOURNAL_AFFINITY_BODY εμφανίζεται ο οδηγός ρύθμισής του. Στη σελίδα General, ενεργοποιούμε την επιλογή **Delete previous content**.

Στο σημείο αυτό η ροή εξόρυξης είναι ρυθμισμένη να κατευθύνει τους κανόνες συσχετίσεων του παραγόμενου μοντέλου σε δύο πίνακες: JOURNAL_AFFINITY και JOURNAL_AFFINITY_BODY.

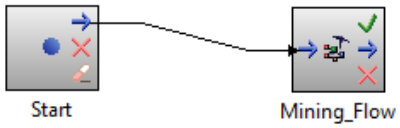
3. Δημιουργία ροής ελέγχου

Στη συνέχεια θα δημιουργήσουμε μία ροή ελέγχου (control flow) η οποία θα είναι υπεύθυνη να εκτελεί τη ροή εξόρυξης.

Μία ροή ελέγχου στο Design Studio μοντελοποιεί τη λογική επεξεργασίας διαφορετικών δραστηριοτήτων (χειριστών). Για τη δημιουργία μίας ροής ελέγχου τοποθετούμε χειριστές σε έναν άδειο καμβά, ορίζουμε τις ιδιότητές τους και τους συνδέουμε μεταξύ τους.

Η διαδικασία είναι απλή:

1. Δημιουργία κενής ροής ελέγχου:
   - Στον Data Project Explorer, δεξί κλικ στον φάκελο **Control Flows** του έργου **HEAL-Link Data Warehousing Project** και επιλογή **New > Control Flow**.
   - Στο παράθυρο δημιουργίας νέας ροής ελέγχου, εισάγουμε ένα όνομα για τη ροή ελέγχου: *HL_control_flow*.
   - Πατάμε το κουμπί **Finish**.
2. Προσθήκη ενός χειριστή ροής εξόρυξης (mining flow operator):
   - Επιλέγουμε τον χειριστή **Mining Flow** από την ομάδα Common Operators της παλέτας χειριστών, και τον τοποθετούμε στον καμβά.
3. Ρύθμιση του χειριστή Mining Flow:
   - Συνδέουμε την πόρτα **Start** του χειριστή Start στην πόρτα εισόδου **Input** του χειριστή Mining Flow.
   - Κάνουμε δεξί κλικ στον χειριστή Mining Flow και επιλέγουμε **Show Properties View**.
   - Στη σελίδα General, κάνουμε κλικ στο κουμπί με τα αποσιωπητικά (...).
   - Στο παράθυρο Mining Flow, επιλέγουμε τη ροή εξόρυξης **GUEST_JOURNAL_AFFINITIES** και πατάμε **OK**.



Σχήμα 7.3. Η ροή ελέγχου

Το επόμενο βήμα είναι η προετοιμασία μίας εφαρμογής data warehouse η οποία θα εγκατασταθεί στον Infosphere Warehouse Server και θα είναι υπεύθυνη για την εκτέλεση της ροής ελέγχου ανά τακτά χρονικά διαστήματα, έτσι ώστε οι πίνακες JOURNAL_AFFINITY και JOURNAL_AFFINITY_BODY να παραμένουν ενημερωμένοι.

4. Προετοιμασία της εφαρμογής εξόρυξης για εγκατάσταση στον Infosphere Warehouse Server

Η προετοιμασία της εφαρμογής εξόρυξης γίνεται με τη δημιουργία ενός συμπιεσμένου πακέτου (data warehouse application package) το οποίο θα περιλαμβάνει τις ροές εξόρυξης και ελέγχου. Το πακέτο αυτό εγκαθίσταται στον Infosphere Warehouse Server.

Η διαδικασία που ακολουθούμε έχει ως εξής:

1. Εκκίνηση του οδηγού Data Warehousing Application Deployment Preparation:
   - Στον Data Project Explorer, δεξί κλικ στο έργο **HEAL-Link Data Warehousing Project** και επιλογή **New > Data Warehousing Application**.
2. Συμπληρώνουμε τον οδηγό ακολουθώντας τα βήματα του παρακάτω πίνακα:

Πίνακας 7.5. Βήματα δημιουργίας νέας εφαρμογής Data Warehousing

Σελίδα	Βήματα
Project Selection	Επιλογή του έργου **HEAL-Link Data Warehousing Project.**
Application Profile	a. Στο πεδίο Profile Name, εισάγουμε το όνομα της εφαρμογής: `HL application`

Σελίδα	Βήματα
	b. Στο πεδίο Description, περιγράφουμε την εφαρμογή: `Discovers journal affinities`
Control Flow Selection	Επιλέγουμε το directory **/HEAL-Link Data Warehousing Project/ control-flows/HL_control_flow** και πατάμε το κουμπί >.
Resource Profile Management	Αποδοχή προεπιλεγμένων τιμών.
Variable Management	Αποδοχή προεπιλεγμένων τιμών.
Miscellaneous File Selection	Αποδοχή προεπιλεγμένων τιμών.
Saving Application Profile	Αποδοχή προεπιλεγμένων τιμών.
Code Generation	Αποδοχή προεπιλεγμένων τιμών.
Package Generation	Στο πεδίο Zip file directory ορίζουμε την τοποθεσία της εφαρμογής: `C:\workspaces\miningtutorial`

Το παράθυρο με τίτλο Package Generation Status επιβεβαιώνει ότι η αποθήκευση του πακέτου της εφαρμογής ολοκληρώθηκε με επιτυχία.

Στην επόμενη παράγραφο θα εγκαταστήσουμε την εφαρμογή στον Infosphere Warehouse Server και θα την προγραμματίσουμε ώστε να εκτελείται ανά εφτά ημέρες.

5. Εγκατάσταση της εφαρμογής εξόρυξης στον Infosphere Warehouse Server

Για την εγκατάσταση της εφαρμογής θα χρησιμοποιήσουμε την κονσόλα διαχείρισης της πλατφόρμας Infosphere Warehouse.

Με την εγκατάσταση και λειτουργία της εφαρμογής στον server, θα υπάρχει η δυνατότητα ανάλυσης των δεδομένων συναλλαγών του συστήματος HEAL-Link καθώς αυτά θα εμπλουτίζονται με το πέρασμα του χρόνου. Η κανόνες συσχετίσεων που θα παράγονται θα είναι πάντα ενημερωμένοι, επιτρέποντας στο σύστημα παραγωγής συστάσεων να κάνει έγκυρες προτάσεις προς τους χρήστες.

Τα βήματα που θα ακολουθήσουμε έχουν ως εξής:

1. Εκκίνηση της κονσόλας διαχείρισης και προσθήκη σύνδεσης προς τη βάση δεδομένων

2. Εγκατάσταση της εφαρμογής εξόρυξης

3. Δημιουργία προγράμματος για την περιοδική εκτέλεση της εφαρμογής εξόρυξης

5.1. Εκκίνηση της κονσόλας διαχείρισης και προσθήκη σύνδεσης προς τη βάση δεδομένων

Πριν την εγκατάσταση της εφαρμογής εξόρυξης θα πρέπει να προσθέσουμε μία νέα σύνδεση προς τη βάση δεδομένων.

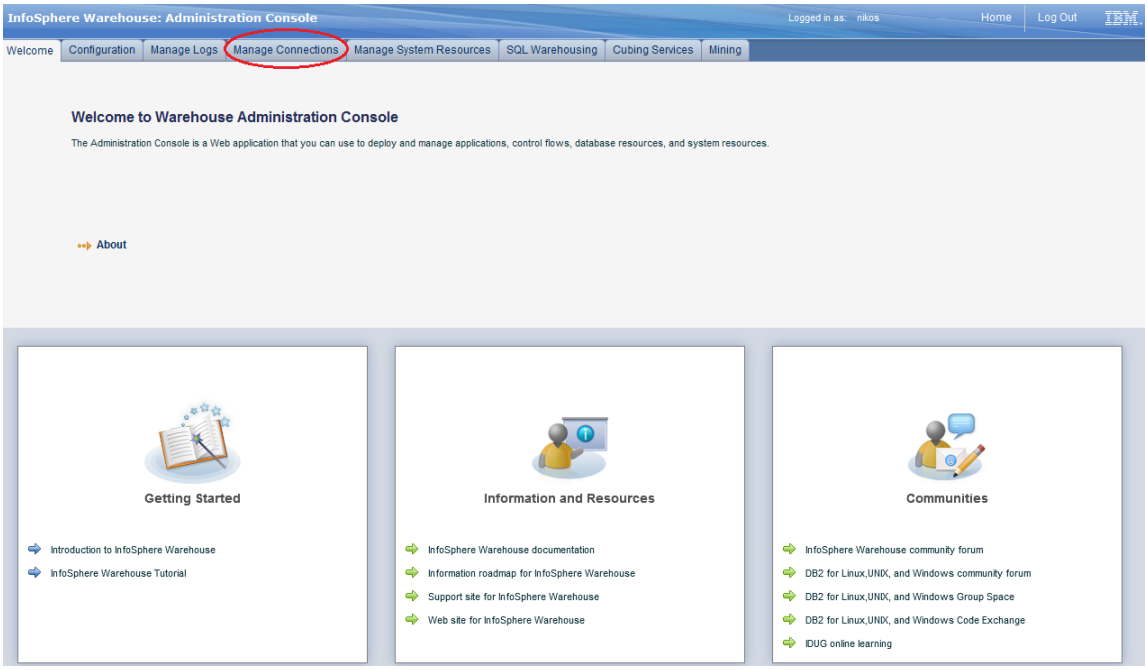
1. Εκκίνηση της κονσόλας διαχείρισης:

   - Επιλέγουμε κατά σειρά **Start > Programs > IBM InfoSphere Warehouse > ISWCOPY01 > Warehouse Administration Console and Workload Manager.**

2. Στη σελίδα εισόδου (Log In), εισάγουμε το όνομα χρήστη της DB2 και τον κωδικό πρόσβασης.

3. Στην περιοχή My Tools της σελίδας καλωσορίσματος, επιλέγουμε **Warehouse Administration Console.**

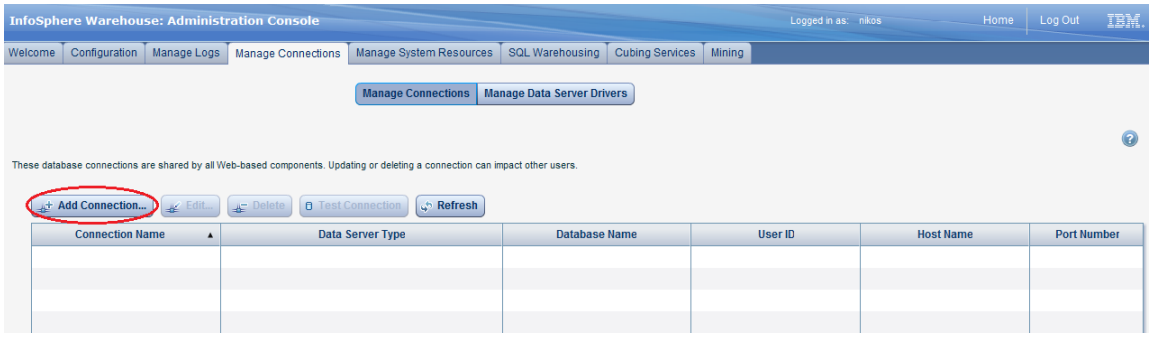
4. Κάνουμε κλικ στην ετικέτα **Manage Connections.**



Σχήμα 7.4. Warehouse Administration Console

5. Εάν δεν υπάρχει σύνδεση προς τη βάση δεδομένων HLDB, προσθέτουμε μία νέα σύνδεση.

   - Κάνουμε κλικ στο κουμπί **Add Connection.**



Σχήμα 7.5. Manage Connections

   - Στο παράθυρο Add Connection, συμπληρώνουμε τα απαραίτητα πεδία.

Πίνακας 7.6. Οι τιμές των πεδίων για την προσθήκη νέας σύνδεσης προς τη βάση δεδομένων HLDB

Πεδίο	Τιμή
Connection name	HLDB
Data server type	**DB2 for Linux, UNIX, and Windows**
Database name	HLDB
Host name	Το όνομα του διακομιστή που φιλοξενεί την DB2 (π.χ. localhost)
Port number	Η πόρτα επικοινωνίας της DB2 (π.χ. 50000)
Encryption method	**Clear text password**
User ID	Όνομα χρήστη της DB2
Password	Κωδικός πρόσβασης του χρήστη

- Στη σελίδα Permissions αποδεχόμαστε τις προεπιλεγμένες τιμές.

6. Έλεγχος της σύνδεσης προς τη βάση δεδομένων:
   - Επιλέγουμε τη σύνδεση προς τη βάση HLDB.
   - Κάνουμε κλικ στο κουμπί **Test Connection.**
   - Στο παράθυρο επιβεβαίωσης του ελέγχου, πατάμε το OK.

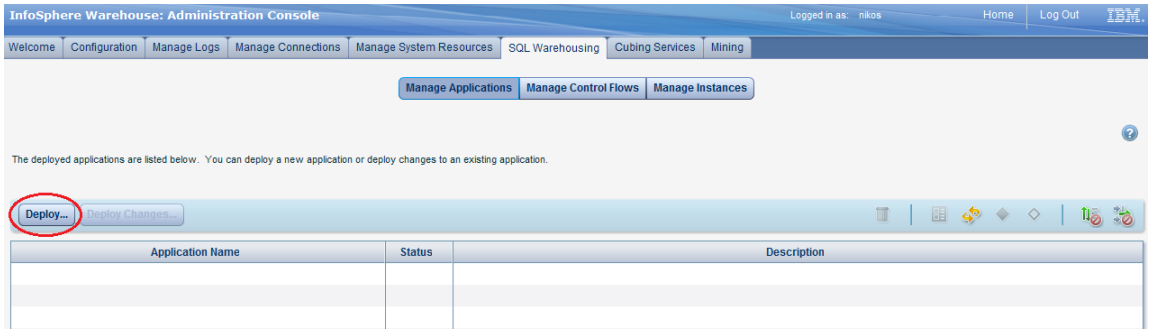
   Το παράθυρο επιτυχίας επιβεβαιώνει ότι η σύνδεση ρυθμίστηκε σωστά.

Στη συνέχεια θα χρησιμοποιήσουμε την κονσόλα διαχείρισης για να εγκαταστήσουμε την εφαρμογή εξόρυξης στον Infosphere Warehouse Server.

5.2. Εγκατάσταση της εφαρμογής εξόρυξης

Για να εγκαταστήσουμε την εφαρμογή εξόρυξης στον server:

1. Κάνουμε κλικ στην ετικέτα **SQL Warehousing.**
2. Κάνουμε κλικ στο κουμπί **Deploy** και ακολουθούμε τον οδηγό Deploy an Application.



Σχήμα 7.6. Εγκατάσταση της εφαρμογής εξόρυξης

Πίνακας 7.7. Οδηγός εγκατάστασης της εφαρμογής εξόρυξης

Σελίδα	Βήματα
Specify the Application File	a. Επιλέγουμε **Zip file location on client.** b. Ορίζουμε την τοποθεσία στην οποία αποθηκεύσαμε την εφαρμογή εξόρυξης: `C:\workspaces\miningtutorial`

Application Parameters	a. Στο πεδίο **Application name**, ονομάζουμε την εφαρμογή. Για παράδειγμα:
	`hl-weekly-update`
	b. Στο πεδίο **Home directory**, ορίζουμε για την εφαρμογή έναν κατάλογο εργασίας (working directory). Για παράδειγμα:
	`C:\hl-working`
	c. Στο πεδίο **Log directory**, ορίζουμε έναν κατάλογο για τα αρχεία καταγραφών της εφαρμογής (logs). Για παράδειγμα:
	`C:\hl-working\logs`

Data Source Mappings	Στη στήλη **Mapped Resource** της γραμμής HLDB, επιλέγουμε **HLDB.**

Σελίδα	Βήματα
System Resource Mappings	Δε χρειάζονται ρυθμίσεις σε αυτή τη σελίδα.
Variable Mappings	Δε χρειάζονται ρυθμίσεις σε αυτή τη σελίδα.
Ready to Deploy	Δε χρειάζονται ρυθμίσεις σε αυτή τη σελίδα.

Η εφαρμογή εξόρυξης είναι πλέον διαθέσιμη στον Infosphere Warehouse Server. Στη συνέχεια, θα δημιουργήσουμε ένα πρόγραμμα με βάση το οποίο θα εκτελείται περιοδικά.

5.3. Δημιουργία προγράμματος για την περιοδική εκτέλεση της εφαρμογής εξόρυξης

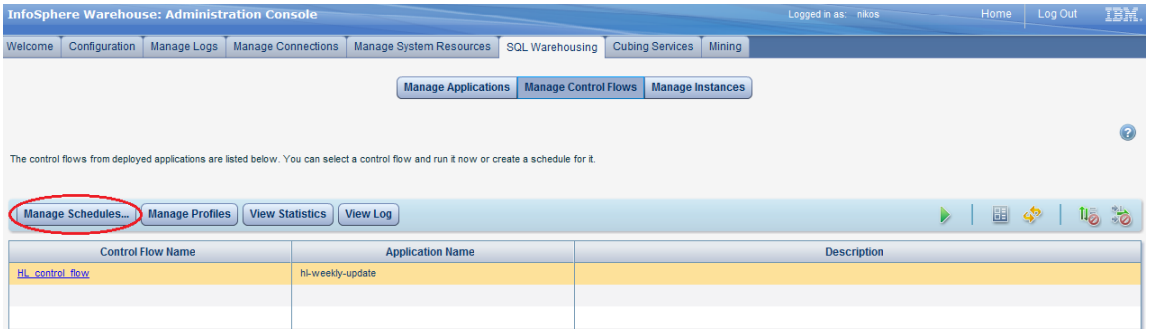
Λόγω της συνεχούς εισροής δεδομένων συναλλαγών στη βάση δεδομένων HLDB, οι συσχετίσεις μεταξύ των αντικειμένων της βάσης ενδέχεται να αλλάζουν. Για να παραμένουν οι κανόνες συσχετίσεων ενημερωμένοι, θα πρέπει να δημιουργήσουμε ένα πρόγραμμα στον server με βάση το οποίο θα επαναλαμβάνεται η εκτέλεση της εφαρμογής εξόρυξης.

Για τη δημιουργία του προγράμματος εκτελούμε:

1. Στη σελίδα SQL Warehousing, κάνουμε κλικ στο κουμπί **Manage Control Flows.**

2. Δημιουργούμε ένα νέο πρόγραμμα για την εφαρμογή εξόρυξης.

   - Επιλέγουμε **HL control flow** και κάνουμε κλικ στο κουμπί **Manage Schedules.**



Σχήμα 7.7. Διαχείριση προγραμμάτων

   - Στο παράθυρο Schedules for Control Flows κάνουμε κλικ στο κουμπί **Create.**

   - Συμπληρώνουμε τον οδηγό δημιουργίας νέου προγράμματος.

Πίνακας 7.8. Οδηγός δημιουργίας νέου προγράμματος

Σελίδα	Βήματα
Schedule Name	Στο πεδίο **Schedule Name** εισάγουμε το όνομα του προγράμματος. Για παράδειγμα: `HL_schedule`
Update Variables	Δε χρειάζονται ρυθμίσεις σε αυτή τη σελίδα.
Schedule Timing	a. Επιλέγουμε **Repeats** και αλλάζουμε τη συχνότητα από **Daily** σε **Weekly.** b. Επιλέγουμε **Indefinitely.** c. Ενεργοποιούμε την επιλογή **On the following days** και επιλέγουμε **Sunday.**

Πλέον, η εφαρμογή εξόρυξης θα εκτελείται ανά εφτά ημέρες ενημερώνοντας τους κανόνες συσχετίσεων με βάση τις τελευταίες αλλαγές στα δεδομένα συναλλαγών.

6. Παραγωγή συστάσεων

Το μοντέλο συσχετίσεων και οι κανόνες που περιλαμβάνει είναι πλέον διαθέσιμοι σε τρίτες εφαρμογές. Ένα σύστημα παραγωγής συστάσεων μπορεί πολύ εύκολα να ζητήσει πληροφορίες από τους πίνακες JOURNAL_AFFINITY και JOURNAL_AFFINITY_BODY οι οποίοι περιλαμβάνουν τους κανόνες συσχετίσεων. Αν, για παράδειγμα, ένας επισκέπτης εισέλθει στον διαδικτυακό τόπο του HEAL-Link και στοχοποιήσει τα συγγράμματα με κωδικούς 44, 45 και 48, το σύστημα παραγωγής συστάσεων θα μπορέσει να θέσει στη βάση δεδομένων το ακόλουθο ερώτημα για να ανακαλύψει ποια άλλα συγγράμματα είναι πιθανό να ενδιαφέρουν τον χρήστη:

```
-- Select ITEMs and ITEMNAMEs from applicable rules.
SELECT DISTINCT HEAD AS ITEM, HEADNAME AS ITEMNAME
FROM DB2ADMIN.JOURNAL_AFFINITY
WHERE
   BODYID IN (
      -- Select BODYIDs that apply to the selected items (44, 45, 48) by
      -- excluding irrelevant rules.
      SELECT DISTINCT BODYID
      FROM DB2ADMIN.JOURNAL_AFFINITY_BODY
      WHERE
         BODYID NOT IN (
            -- Select BODYIDs of records that do not include the items.
            -- These BODYIDs indicate rules that are not relevant.
            SELECT BODYID
            FROM DB2ADMIN.JOURNAL_AFFINITY_BODY
            WHERE
               ITEM NOT IN (44,45,48)
         )
   )
   -- Rules must satisfy these quality metrics
   AND SUPPORT > 0.001
   AND CONFIDENCE > 0.3
   AND LIFT > 5
```

Το παραπάνω ερώτημα εντοπίζει συναφείς κανόνες συσχετίσεων οι οποίοι ικανοποιούν τους περιορισμούς ποιότητας που θέτουμε (support, confidence και lift). Με βάση αυτούς τους κανόνες, μπορούμε να προτείνουμε στον χρήστη συγγράμματα που είναι πιθανό να τον ενδιαφέρουν.

Αρχικά, ανακτούμε από τον πίνακα JOURNAL_AFFINITY_BODY τα BODYIDs των εγγραφών που δεν περιέχουν κανένα από τα αντικείμενα της ομάδας αντικειμένων που επέλεξε ο χρήστης (44, 45, 48). Αυτά τα BODYIDs σηματοδοτούν τους κανόνες που θα αποκλείσουμε γιατί δεν μας προσφέρουν καμία χρήσιμη πληροφορία, αφού δε σχετίζονται με τις επιλογές του χρήστη. Στη συνέχεια, επιλέγουμε από τον πίνακα JOURNAL_AFFINITY_BODY εκείνα τα BODYIDs που δεν ανήκουν στην ομάδα που ανέκτησε το προηγούμενο υποερώτημα (δηλαδή στην ομάδα των BODYIDs που σηματοδοτούν τους μη σχετιζόμενους κανόνες). Με άλλα λόγια, ανακτούμε εκείνα τα BODYIDs τα οποία περιέχουν τουλάχιστον ένα αντικείμενο από αυτά που επέλεξε ο χρήστης. Τέλος, επιλέγουμε από τον πίνακα JOURNAL_AFFINITY τις στήλες HEAD και HEADNAME (δηλαδή τις τελικές προτάσεις) όλων των εγγραφών που έχουν BODYID ίσο με κάποια από τις τιμές που επέστρεψε το τελευταίο υποερώτημα (δηλαδή τα BODYIDs που σχετίζονται με τις επιλογές του χρήστη). Οι κανόνες που θα επιστραφούν θα πρέπει να έχουν support μεγαλύτερο από 0.001, confidence μεγαλύτερο από 0.3 και lift μεγαλύτερο από 5.

Κεφάλαιο 8. Σύνοψη και περαιτέρω ανάπτυξη

Έχοντας ακολουθήσει όλες της φάσεις της διεργασίας εξόρυξης πληροφορίας από τα δεδομένα χρήσης του HEAL-Link portal, προέκυψε μία αναλυτική μελέτη ανακάλυψης κρυμμένων συσχετίσεων στα δεδομένα συναλλαγών της διαδικτυακής πύλης. Ξεκινήσαμε με την κατανόηση της επιχείρησης και των επιχειρηματικών στόχων, περιγράψαμε την τεχνική εξόρυξης των κανόνων συσχετίσεων, εισαγάγαμε και εξερευνήσαμε τα δεδομένα, προετοιμάσαμε τα δεδομένα για τον αλγόριθμο εξόρυξης, αναπτύξαμε και αξιολογήσαμε τα μοντέλα κανόνων συσχετίσεων. Τέλος, παρουσιάσαμε έναν τρόπο με τον οποίο μπορεί να αναπτυχθεί μία λειτουργική λύση ενός συστήματος παραγωγής συστάσεων για την εξυπηρέτηση των χρηστών του HEAL-Link portal.

Τα προϊόντα της εργασίας θα μπορούσαν να μελετηθούν από την ομάδα διαχείρισης του διαδικτυακού τόπου, υποστηρίζοντάς τους στη λήψη αποφάσεων για τη βελτίωση των προσφερόμενων υπηρεσιών προς τους χρήστες.

Ως συνέχεια της εργασίας, θα μπορούσε να οριστεί ο σχεδιασμός και η ανάπτυξη ενός υποσυστήματος λογισμικού το οποίο θα αξιοποιεί την εφαρμογή εξόρυξης παράγοντας συστάσεις προς τους χρήστες της πύλης σε πραγματικό χρόνο. Το υποσύστημα αυτό θα μπορούσε να ενσωματωθεί πιλοτικά στην διαδικτυακή πύλη του HEAL-Link. Μία άλλη οδός συνέχειας της παρούσης εργασίας, θα ήταν η εφαρμογή εναλλακτικών τεχνικών εξόρυξης στα δεδομένα συναλλαγών του HEAL-Link, όπως για παράδειγμα συσταδοποίησης ή/και συνδυασμού συσταδοποίησης με κανόνες συσχετίσεων. Ταυτόχρονα, θα μπορούσαν να μελετηθούν περαιτέρω οι δυνατότητες της πλατφόρμας Infosphere Warehouse και να εφαρμοστεί στην πράξη η τεχνολογία Scoring του Intelligent Miner (μία τεχνολογία εφαρμογής των μοντέλων εξόρυξης σε νέα δεδομένα).

Βιβλιογραφία

Χαλκίδη, Μ., και Βαζιργιάννης, Μ. (2005), Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, τυπωθήτω - ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ

Elmasri, R., and Navathe, S.B. (2005), Θεμελιώδεις Αρχές Συστημάτων Βάσεων Δεδομένων, ΔΙΑΥΛΟΣ

Ramakrishnan, R., and Gehrke, J. (2002), Συστήματα Διαχείρισης Βάσεων Δεδομένων, 2nd Edition, Τόμοι Α και Β, ΤΖΙΟΛΑ

Ballard et al., (2007), Dynamic Warehousing: Data Mining Made Easy, IBM Redbooks

Ballard et al., (2006), Leveraging DB2 Data Warehouse Edition for Business Intelligence, IBM Redbooks

Garcia Bordes M. Eugenia, (2006), Mining your Business in Retail with IBM DB2 Intelligent Miner, IBM Corporation

Chapman et al., (2000), CRISP-DM 1.0 Step-by-step data mining guide, SPSS

Οδηγός χρήσης λογισμικού

Το πρόγραμμα TimestampConverter υποβλήθηκε σε ηλεκτρονική μορφή μαζί με την πτυχιακή εργασία. Συγκεκριμένα, υποβλήθηκαν τα ακόλουθα δύο συμπιεσμένα αρχεία:

1. NetBeans_Project_TimestampConverter.rar

2. TimestampConverter_v001.rar

Το πρώτο αρχείο περιλαμβάνει το NetBeans project του TimestampConverter. Ο TimestampConverter αναπτύχθηκε στο περιβάλλον ανάπτυξης NetBeans V6.9. Για την μελέτη του κώδικα του προγράμματος θα πρέπει να γίνει εξαγωγή του project από το συμπιεσμένο αρχείο και φόρτωσή του στο NetBeans IDE.

Το δεύτερο αρχείο περιλαμβάνει το τελευταίο build του προγράμματος. Πληροφορίες για την εγκατάσταση και χρήση του υπάρχουν στο README.html εντός του TimestampConverter_v001.rar.