ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Θέμα: «Market Basket Analysis»

Στυλιανός Μητσοτάκης Α.Μ 13826

Επιβλέπων καθηγητής: Βασίλειος Ταμπακάς

ΑΝΤΙΡΡΙΟ 2018

Περιεχόμενα

Κεφάλαιο 1ο

1.1 Εισαγωγή

Είναι βέβαιο ότι ζούμε στην κοινωνία της πληροφορίας, όπου η μετατροπή των δεδομένων σε πληροφορία απαιτείται να οδηγεί στη μετατροπή της πληροφορίας σε γνώση. Μια από τις πιο προκλητικές εργασίες της εποχής μας είναι η ανακάλυψη προτύπων, τάσεων και ανωμαλιών σε τεράστια σύνολα δεδομένων, καθώς και η σύνοψή τους μέσω απλών και εύχρηστων μοντέλων. Τα προβλήματα της εξόρυξης δεδομένων ως τεχνικές έχουν προσεγγιστεί από ετερόκλητα επιστημονικά πεδία όπως της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας της πληροφορίας και των υπολογιστικών διαδικασιών, έχει δημιουργήσει μια νέα επιστήμη με δυναμικά εργαλεία, η οποία καλείται «Εξόρυξη Δεδομένων» και είναι μέρος της διαδικασίας «Ανακάλυψης Γνώσης από Βάσεις Δεδομένων».

Η σύγκλιση της προόδου υπολογιστικών συστημάτων και της εξέλιξης στην επικοινωνία έχει οδηγήσει στην δημιουργία μιας κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό που συγκεντρώνεται καταγράφεται διαρκώς, με αποτέλεσμα τη δημιουργία τεράστιων βάσεων δεδομένων. Το ζήτημα λοιπόν που προκύπτει, είναι εάν μπορούμε να διαχειριστούμε αυτές τις βάσεις δεδομένων. Όλα αυτά τα θέματα προκάλεσαν το ενδιαφέρον και οδήγησαν στη διαδικασία της Εξόρυξης Δεδομένων (Data Mining). Πρόκειται για μία σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς και ετερόκλητους κλάδους όπως οι: οικονομία, βιοστατιστική, δημογραφία, μετεωρολογία και γεωλογία. Υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων.

Ωστόσο, ο ακόλουθος ορισμός, θεωρείται αξιόλογος:

«Εξόρυξη Δεδομένων είναι η ανάλυση – συνήθως τεράστιων – παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων». Η δήλωση των σχέσεων και η σύνοψη των στοιχείων στην οποία αναφέρεται ο ορισμός αυτός, συχνά αναφέρεται ως μοντέλο ή πρότυπο.

Οι δυο βασικοί στόχοι της εξόρυξης δεδομένων είναι η εφαρμογή τεχνικών περιγραφής και πρόβλεψης σε μεγάλα σύνολα δεδομένων.

Η πρόβλεψη στοχεύει στον υπολογισμό της μελλοντικής αξίας ή στην πρόβλεψη της συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και οι οποίες βασίζονται στην συμπεριφορά άλλων μεταβλητών. Η περιγραφή επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με ένα κατανοητό και αξιοποιήσιμο τρόπο. Η σημαντικότητα της πρόβλεψης και της περιγραφής διαφέρει ανάλογα με τις εφαρμογές εξόρυξης δεδομένων. Ωστόσο ως προς την εξόρυξη γνώσης η περιγραφή τείνει να είναι περισσότερο σημαντική από την πρόβλεψη σε αντίθεση με την αναγνώριση προτύπων και την εφαρμογή μηχανικής μάθησης για τις οποίες η πρόβλεψη είναι πιο σημαντική.
Ένας αριθμός μεθόδων εξόρυξης δεδομένων έχουν προταθεί για να ικανοποιούν τις απαιτήσεις διαφορετικών εφαρμογών. Ωστόσο όλες επιτυγχάνουν μια ομάδα από διεργασίες εξόρυξης δεδομένων για να προσδιορίσουν και να περιγράψουν ενδιαφέροντα πρότυπα γνώσης που έχουν αντληθεί από ένα σύνολο δεδομένων. Θα αναφερθούμε στη συνέχεια της διπλωματικής σε αυτές τις μεθόδους αναλυτικά.

1.2 Data Mining

Data mining (εξόρυξη δεδομένων) είναι η διαδικασία ψαξίματος σε μεγάλο όγκο δεδομένων, με σκοπό την περισυλλογή πληροφορίας. Το data mining έχει περιγραφεί σαν: «η μη τετριμμένη εξαγωγή υπονοούμενης, ενδεχομένως χρήσιμης και μέχρι στιγμής άγνωστης πληροφορίας από διάφορα δεδομένα» και «η επιστήμη της εξαγωγής χρήσιμης πληροφορίας από μεγάλα datasets ή βάσεις δεδομένων». Σε σχέση με το ERP (Enterprise Resource Planning), το data mining ορίζεται σαν η στατιστική και λογική ανάλυση μεγάλου συνόλου δεδομένων που προέκυψαν από συναλλαγές, με σκοπό την εύρεση patterns.

Το data mining λογισμικό είναι ένα από τα μερικά αναλυτικά εργαλεία (analytical tools) που υπάρχουν για ανάλυση δεδομένων. Επιτρέπει στους χρήστες να αναλύουν δεδομένα από πολλές διαφορετικές διαστάσεις και γωνίες, να τα κατηγοριοποιούν και να συνοψίσουν τις συσχετίσεις που υπάρχουν. Πιο τεχνικά, το data mining ορίζεται σαν η διαδικασία της εύρεσης συσχετίσεων και μοτίβων (patterns) μεταξύ δεκάδων πεδίων μεγάλων βάσεων δεδομένων .

Ανέκαθεν διάφοροι αναλυτές ασχολούνται με την διαδικασία της εξαγωγής χρήσιμης πληροφορίας από καταγεγραμμένα δεδομένα. Στις σύγχρονες όμως επιχειρήσεις και επιστήμες, ο αυξημένος όγκος των δεδομένων απαιτεί το ψάξιμο αυτό να γίνεται με χρήση υπολογιστικών προσεγγίσεων. Πράγματι, στην σύγχρονη εποχή, ο όγκος των δεδομένων αυτών, τα οποία δεδομένα ονομάζονται datasets, έχει αυξηθεί τόσο σε μέγεθος, όσο και σε πολυπλοκότητα. Το αποτέλεσμα είναι οι μέθοδοι απ`ευθείας ανάλυσης δεδομένων να δίνουν την θέση τους σε αυτόματες μεθόδους ανάλυσης δεδομένων, οι οποίες χρησιμοποιούν πιο πολύπλοκα και εξεζητημένα εργαλεία.

Από την άλλη, οι αναπτυσσόμενες τεχνολογίες των υπολογιστών, δικτύων και αισθητήρων έχουν μετάτρεψε την περισυλλογή και οργάνωση δεδομένων σε μια αρκετά εύκολη διαδικασία. Το ζητούμενο όμως δεν είναι μόνο η περισυλλογή των δεδομένων αυτών. Τουναντίον, τα περισυλλεγμένα δεδομένα πρέπει να μετατραπούν σε κατάλληλη πληροφορία και γνώση, για να μπορούν να καταστούν χρήσιμα. Έτσι, μπορούμε να πούμε ότι το data mining αποτελεί την διαδικασία φαρμογής υπολογιστικών μεθοδολογιών, συμπεριλαμβανομένων και καινούργιων τεχνικών, με σκοπό την ανακάλυψη γνώσης.

Παρόλο που το data mining είναι ένας σχετικά καινούργιος όρος, εντούτοις η τεχνολογία δεν είναι καινούργια. Διάφορες επιχειρήσεις ανέκαθεν χρησιμοποιούσαν ισχυρούς υπολογιστές για να «κοσκινίσουν» τεράστιους όγκους δεδομένων, όπως για παράδειγμα δεδομένα που προκύπτουν από τον σαρωτή προϊόντων (optical scanner ή barcode reader) κάποιας υπεραγοράς, με σκοπό την παραγωγή διαφόρων αναφορών σχετικών με την αγορά. Εξάλλου, η συνεχιζόμενη και αυξανόμενη εμφάνιση καινοτομιών σχετικών με την απόδοση των μοντέρνων υπολογιστικών συστημάτων, την απόδοση και χωρητικότητα των συστημάτων αποθήκευσης αλλά και την παραγωγή λογισμικού σχετικού με την στατιστική ανάλυση, αυξάνουν δραματικά την ακρίβεια και την χρησιμότητα της ανάλυσης των δεδομένων αυτών, μειώνοντας παράλληλα το κόστος.

Ο όρος data mining χρησιμοποιείται συχνά αναφερόμενος σε δύο διαφορετικές διαδικασίες: την ανακάλυψη γνώσης και την πρόβλεψη. Η ανακάλυψη γνώσης παρέχει

ρητή πληροφορία η οποία έχει αναγνώσιμη μορφή και μπορεί να γίνει κατανοητή από κάποιον χρήστη. Η πρόβλεψη παρέχει προβλέψεις για μελλοντικά συμβάντα. Σε κάποιες περιπτώσεις μπορεί να είναι διαφανείς και αναγνώσιμη, π.χ. σε συστήματα βασισμένα σε κανόνες (rule based systems), ενώ σε άλλες περιπτώσεις μπορεί να είναι αδιαφανής, όπως σε νευρωνικά δίκτυα (neural networks). Εφόσον το data mining βασίζεται στην χρήση δεδομένων του πραγματικού κόσμου, μπορεί σε κάποιες περιπτώσεις τα δεδομένα αυτά να είναι εξαιρετικά ευαίσθητα και να έχουν άγνωστες αλληλοσυσχετίσεις. Μια αναπόφευκτη αδυναμία του data mining είναι ότι κρίσιμα δεδομένα που μπορεί να έχουν αλληλοσυσχετίσεις δεν παρατηρούνται ποτέ.Ανεξάρτητα από τις προσπάθειες αυτές, υπάρχουν και ελεύθερα «ανοικτού» κώδικα συστήματα όπως το RapidMiner και το Weka, τα οποία αποτελούν το ανεπίσημο standard για τον ορισμό των data mining διαδικασιών. Πιο συγκεκριμένα για το Weka θα δούμε σε επόμενο κεφάλαιο.

1.2.1 Δεδομένα, πληροφορία και γνώση

Δεδομένα

Σαν δεδομένα μπορούμε να ορίσουμε οτιδήποτε (αριθμούς, κείμενο κ.τ.λ.) μπορεί να επεξεργαστεί κάποιο υπολογιστικό σύστημα. Σήμερα, διάφοροι οργανισμοί συγκεντρώνουν κολοσσιαίες ποσότητες δεδομένων σε πολλές διαφορετικές τυποποιήσεις (formats) και βάσεις δεδομένων. Τα δεδομένα αυτά συμπεριλαμβάνουν:

- Επιχειρησιακά δεδομένα ή δεδομένα συναλλαγών όπως πωλήσεις, απογραφές, μισθολόγια, λογιστικά.
- Μη επιχειρησιακά δεδομένα, όπως δεδομένα προβλέψεων και μακροοικο-νομικά δεδομένα.
- Μεταδεδομένα (δεδομένα που περιγράφουν δεδομένα) όπως η σχεδίαση λογικών βάσεων δεδομένων και προσδιορισμοί δεδομένων λεξικών.

Πληροφορία

Τα μοτίβα (patterns), οι συσχετίσεις (associations) και οι συνάφειες (relationships) μεταξύ όλων αυτών των δεδομένων, μπορούν να παρέχουν πληροφορία. Για παράδειγμα, η ανάλυση δεδομένων από συναλλαγές που έγιναν σε ένα σημείο πώλησης προϊόντων (όπως μια υπεραγορά), μπορεί να παράγει πληροφορία σχετικά με το ποια προϊόντα πωλούνται, πότε (χρονικά διαστήματα) και πώς (με ποια άλλα προϊόντα πωλούνται μαζί κ.ο.κ.).

Γνώση

Η πληροφορία με την σειρά της μπορεί να μετατραπεί σε γνώση, η οποία να αφορά μελλοντικές τάσεις της αγοράς. Στο παράδειγμα της υπεραγοράς πιο πάνω, η πληροφορία που κερδίζεται με την ανάλυση των δεδομένων από τις συναλλαγές, μπορεί να βοηθήσει στον προσδιορισμό της αγοραστικής συμπεριφοράς των πελατών, δηλαδή να γίνει γνωστή και να εκμεταλλευτεί η αγοραστική τάση του κοινού. Γνωρίζοντας την τάση αυτή, είναι εφικτός ο προσδιορισμός των προϊόντων εκείνων που είναι καλύτερα και πρέπει να προωθούνται περισσότερο.

1.2.2 Χρήσεις του data mining

Το data mining χρησιμοποιείται ευρέως σε διάφορους τομείς, όπως στης επιχειρήσεις, επιστήμες, μηχανική κ.α.

Επιχειρήσεις

Σήμερα, το data mining χρησιμοποιείται κυρίως από διάφορες επιχειρήσεις που ασχολούνται άμεσα με τους καταναλωτές, επιχειρήσεις πώλησης προϊόντων, επιχειρήσεις που ασχολούνται με την οικονομία, τις επικοινωνίες αλλά και επιχειρήσεις που ασχολούνται με το marketing. Το data mining βοηθά τις επιχειρήσεις αυτές στον καθορισμό συναφειών μεταξύ εσωτερικών παραγόντων όπως οι τιμές, η τοποθέτηση των προϊόντων και η ικανότητα του προσωπικού, αλλά και μεταξύ εξωτερικών παραγόντων όπως οι οικονομικοί δείκτες και ο ανταγωνισμός. Βοηθά στην μελέτη του αντίκτυπου στις πωλήσεις από διάφορες καταστάσεις, το ποσοστό ικανοποίησης των πελατών και τον καθορισμό του κέρδους, ενώ επιβάλλει την εις βάθος μελέτη και επεξεργασία των δεδομένων των συναλλαγών.

Με την χρήση data mining τεχνικών, ένας πωλητής μπορεί να χρησιμοποιήσει δεδομένα συναλλαγών πελατών του για να μάθει το αγοραστικό ιστορικό τους και στην συνέχεια να το εκμεταλλευτεί, προωθώντας σε αυτούς με κατάλληλο τρόπο διάφορα προϊόντα. Επιπλέον, με την εξόρυξη δημογραφικών δεδομένων από διάφορες φόρμες σχολίων και κάρτες εγγύησης διαφόρων πελατών, ένας πωλητής θα μπορούσε να δημιουργήσει δελεαστικές προσφορές που να έχουν σαν στόχο συγκεκριμένη μερίδα πελατών.

Για παράδειγμα, καταστήματα ενοικίασης ταινιών εφαρμόζουν τεχνικές εξόρυξης σε βάσεις δεδομένων που διατηρούν με το ιστορικό των ενοικιάσεων των πελατών τους, με σκοπό την παραγωγή κατάλληλης πληροφορίας, ώστε να μπορούν να προτείνουν στους πελάτες τους διάφορες ταινίες. Ακόμα, εταιρίες πιστωτικών καρτών προτείνουν προϊόντα στους πελάτες τους, ανάλογα με τις προηγούμενες αγορές που οι τελευταίοι έχουν κάνει.

Φυσικά, η πληροφορία για το ποια προϊόντα ενδιαφέρουν κάποιο πελάτη, προκύπτει με ανάλυση των προηγούμενών τους αγορών, με τη χρήση data mining τεχνικών.

Επιχειρήσεις πώλησης αγαθών τεράστιων διαστάσεων πραγματοποιούν συνεχώς περισυλλογή δεδομένων συναλλαγών πελατών τους ταυτόχρονα από χιλιάδες καταστήματα από διάφορες χώρες και τα προωθούν (τα δεδομένα αυτά) σε κολοσιαία (πολλών Terabyte) data warehouses. Στην συνέχεια, οι επιχειρήσεις αυτές επιτρέπουν σε χιλιάδες προμηθευτές τους να έχουν πρόσβαση στα δεδομένα και να τα αναλύουν με τεχνικές data mining. Οι προμηθευτές χρησιμοποιούν τα δεδομένα αυτά για να αναγνωρίσουν αγοραστικά μοτίβα πελατών και να ανακαλύψουν νέες ευκαιρίες που σχετίζονται με το εμπόριο διαφόρων προϊόντων.

Το data mining συνεισφέρει επίσης σε εφαρμογές διαχείρισης των σχέσεων με τους πελάτες. Για παράδειγμα, μια επιχείρηση θα μπορούσε να εφαρμόσει data mining τεχνικές για την εύρεση των πελατών με την μεγαλύτερη πιθανότητα να απαντήσουν σε κάποια δική τους προσφορά, παρά την μαζική αποστολή email και ενημερωτικών φυλλαδίων σε αυτούς. Ποιό εξειδικευμένες τεχνικές data mining μπορούν να χρησιμοποιηθούν στο να προβλέπουν και ποιες συγκεκριμένες προσφορές είναι πιθανότερο να ενδιαφέρουν κάποιο πελάτη. Ακόμα, αντί για ένα μοντέλο το οποίο να προβλέπει ποιοι από τους πελάτες θα αποδεχθούν τις προσφορές, μια επιχείρηση μπορεί να εφαρμόσει πολλά διαφορετικά μοντέλα για πολλούς διαφορετικούς χρήστες.

Μια ακόμα εφαρμογή του data mining είναι σε τμήματα ανθρωπίνων πόρων (human-resources departments). Σε τέτοιες περιπτώσεις, το ζητούμενο είναι το να προσδιοριστούν τα χαρακτηριστικά των πιο επιτυχημένων υπαλλήλων μιας επιχείρησης. Τέτοια χαρακτηριστικά μπορεί να είναι για παράδειγμα το πανεπιστήμιο στο οποίο φοίτησαν, ούτως ώστε μελλοντικά να γίνει πρόσληψη περισσότερου ανθρώπινου δυναμικού από το πανεπιστήμιο αυτό.

Το market basket analysis είναι μια σημαντική μέθοδος του data mining, η οποία έχει άμεση σχέση με την εφαρμογή της παρούσας διπλωματικής. Η μέθοδος αυτή χρησιμοποιείται στα τμήματα πωλήσεων προϊόντων, είτε για την επιλογή πελατών με κάποιες συγκεκριμένες προτιμήσεις, ή για την επιλογή προϊόντων με κάποια συγκεκριμένα χαρακτηριστικά. Οι διάφορες προτιμήσεις των πελατών αλλά και τα προϊόντα με τις συγκεκριμένες ιδιότητες είναι δυνατόν να προκύψουν με ανάλυση των αγορών των πελατών. Έτσι, αν ένα κατάστημα εμπορίου ρούχων καταγράφει τις αγορές των πελατών του, τότε με κάποιο σύστημα data mining θα μπορούσε να εξαχθεί πληροφορία της μορφής: «ποιοι πελάτες προτιμούν το μετάξι αντί για το βαμβάκι» ή «τι ποσοστό των αγορών που περιλαμβάνουν μετάξι, περιλαμβάνουν και βαμβάκι». Τα ποιό πάνω προκύπτουν με κατάλληλη ανάλυση των συναλλαγών των πελατών, για εξαγωγή των λεγόμενων κανόνων συσχέτισης (association rules). Για την περίπτωση ενός καταστήματος εμπορίου ρούχων, μπορεί να προκύψουν κανόνες όπως: «το 79% των αγορών που περιλαμβάνουν μετάξι, περιλαμβάνουν και βαμβάκι». Στην περίπτωση βιομηχανίας κατασκευής προϊόντων, ένας κανόνας συσχέτισης θα μπορούσε να είναι: «Το 82% των προϊόντων που έχουν ένα συγκεκριμένο κατασκευαστικό λάθος, θα παρουσιάσουν ένα δεύτερο πρόβλημα σε περίοδο 6 μηνών». Η μέθοδος market basket analysis και η διαδικασία εξαγωγής κανόνων συσχέτισης παρουσιάζονται σε επόμενο κεφάλαιο.

Επιστήμες - Μηχανική

Στην σύγχρονη εποχή,το data mining χρησιμοποιείται ευρέως στους τομείς των επιστημών και της μηχανικής. Πιο συγκεκριμένα, απαντάται στην βιοπληροφορική, την γενετική, τον φαρμακευτικό τομέα, την εκπαίδευση και την ηλεκτρολογία. Στον τομέα της γενετικής, ο στόχος είναι να βρούμε το πώς οι αλλαγές στην αλυσίδα DNA κάποιου ατόμου επηρεάζουν το ρίσκο της ανάπτυξης κοινών ασθενειών, όπως είναι ο καρκίνος. Αυτό είναι πολύ σημαντικό για την βελτίωση της διάγνωσης, αποφυγής και θεραπείας των διαφόρων ασθενειών. Η τεχνική data για την διαδικασία αυτή ονομάζεται multifactor dimensionality reduction.

Στην περιοχή της ηλεκτρολογίας, data mining τεχνικές έχουν χρησιμοποιηθεί ευρέως για επίβλεψη της κατάστασης ηλεκτρολογικού εξοπλισμού υψηλής τάσης. Ο σκοπός της επίβλεψης είναι η λήψη πολύτιμης πληροφορίας που αφορά την κατάσταση της μόνωσης του εξοπλισμού. Τέλος, το data mining εφαρμόζεται στην εκπαιδευτική έρευνα (educational research), όπου χρησιμοποιείται προς μελέτη των παραγόντων που οδηγούν τους μαθητές/φοιτητές σε δραστηριότητες, οι οποίες μειώνουν την μάθηση.

1.3 Κύρια Στοιχεία του Data Mining

Ομαδοποίηση (Συσταδοποίηση) – clustering

Είναι η εργασία καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων. Αυτό που διαφοροποιεί την ομαδοποίηση από την κατηγοριοποίηση είναι ότι η ομαδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες. Στην κατηγοριοποίηση ο πληθυσμός διαιρείται σε κατηγορίες αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία με βάση ένα μοντέλο που αναπτύσσεται μέσα της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων. Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες. Οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Εμείς καθορίζουμε την σημασία που θα έχει κάθε ομάδα από τις ομάδες που προκύπτουν.

Κανόνες συσχέτισης (Association rule mining)

Η εξαγωγή κανόνων συσχέτισης θεωρείται μια από τι σημαντικότερες διεργασίες εξόρυξης δεδομένων. Έχει προσελκύσει ιδιαίτερο ενδιαφέρον καθώς οι κανόνες συσχέτισης παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων.

Κατηγοριοποίηση (Classification)

Αποτελεί μια από τις βασικές εργασίες εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαρίστανται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις καθορισμένες κατηγορίες. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιεί δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί

Πρότυπα ακολουθιών (Sequential patterns)

Είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικά με το χρόνοή άλλες ακολουθίες. Οι περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνεται στα συμβολικά πρότυπα.

Ομοιότητα Χρονολογικών σειρών

Μια χρονολογική σειρά είναι μια ακολουθία ορισμών κάθε ένας από τους οποίους έχει ένα timestamp(ετικέτα χρόνου). Χαρακτηριστικά υποθέτουμε ότι οι διαδοχικοί αριθμοί χωρίζονται από ένα σταθερό χρονικό διάστημα και το πραγματικό timestamp παραλείπεται. Τα δεδομένα μιας χρονολογικής σειράς είναι πανταχού παρόντα. Διαφορετικές φυσικές διαδικασίες παράγουν δεδομένα υπό μορφή χρονολογικών σειρών οι οποίες εμφανίζονται μεταξύ άλλων στον οικονομικό τομέα, στον περιβαλλοντικό τομέα, στην ασφάλεια.

Παλινδρόμηση

Η παλινδρόμηση αναφέρεται στην εκμάθηση μιας λειτουργίας που εκχωρεί τα δεδομένα σε μια μεταβλητή πρόβλεψης η οποία παίρνει τιμές πραγματικές.

Περιληπτική παρουσίαση πληροφορίας

Περιλαμβάνει τη διαδικασία ανεύρεσης μιας συμπαγής περιγραφής για ένα σύνολο δεδομένων. Οι τεχνικές περιληπτικής παρουσίασης της πληροφορίας εφαρμόζεται συχνά στη διαλογική διερευνητική ανάλυση δεδομένων και την αυτοματοποιημένη παραγωγή εκθέσεων.

1.3.1 Διαχωρισμός των μεθόδων εξόρυξης δεδομένων

Τα τέσσερα μέρη ή ομάδες εργασιών της ΕΔ (data mining tasks) είναι τα ακόλουθα:

- Περιγραφική μοντελοποίηση (descriptive modeling)
- Η μοντελοποίηση πρόβλεψης (predictive modeling),
- Η ανάλυση συνάφειας (association analysis),
- Η ανίχνευση παρεκτροπών (anomaly detection).

Ο στόχος ενός μοντέλου περιγραφής είναι να γίνει περιγραφή όλου του συνόλου δεδομένων ή της διαδικασίας που παράγει τα δεδομένα. Η σημαντικότερη εφαρμογή των περιγραφικών μοντέλων είναι η συσταδοποίηση, η οποία επιχειρεί να βρει ομάδες παρατηρήσεων που είναι κοντά μεταξύ τους ως προς τα χαρακτηριστικά που περιλαμβάνουν. Οι μέθοδοι περιγραφής και ειδικά συσταδοποίηση είναι πολύ χρήσιμες σε πελατοκεντρικές επιχειρήσεις που βασίζονται στο CRM (Customer Relationship Management), καθώς έτσι μπορούν να εντοπιστούν ομάδες πελατών που αναμένεται να έχουν όμοια συμπεριφορά.

Η κατασκευή ενός μοντέλου πρόβλεψης στοχεύει στη δυνατότητα πρόγνωσης της τιμής μιας μεταβλητής (απόκριση) μέσα από τις τιμές άλλων μεταβλητών (επεξηγηματικές) που είναι γνωστές. Εάν η μεταβλητή απόκρισης είναι (ή μπορεί να θεωρηθεί) κατηγορική, τότε είμαστε σε θέση να εφαρμόσουμε μια μέθοδο ταξινόμησης. Όμως, αν έχουμε συνεχή απόκριση, τότε προχωράμε σε του καλαθιού αγοράς» (market basket analysis).

Άλλες εφαρμογές πραγματοποιούνται στην προώθηση προϊόντων ή στην τοποθέτησή τους στα ράφια καταστημάτων, στη διαχείριση αποθεμάτων κ.λπ. Στους κανόνες συνάφειας δίνεται και εκτίμηση για το πόσο πιθανό να συμβεί αυτή η σχέση αιτίας – αποτελέσματος.

Τέλος, στην ανίχνευση παρεκτροπών ανήκουν εργασίες εντοπισμού παρατηρήσεων των οποίων τα χαρακτηριστικά διαφέρουν σημαντικά από αυτά του υπόλοιπου συνόλου δεδομένων (έκτροπες παρατηρήσεις ή outliers). Στόχος είναι η υψηλού επιπέδου ανίχνευση πιθανών ανωμαλιών, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης. Ως εφαρμογή μπορούμε να αναφέρουμε τον προσδιορισμό απειλής στην έγκριση δανείων ή πιστωτικών καρτών από μια τράπεζα.

1.3.2 Διαδικασία Ανακάλυψης Γνώσης

Στα πλαίσια της αναζήτησης περισσότερο αποτελεσματικών και δυναμικών εργαλείων διαχείρισης διαφορετικής φύσεως δεδομένων, ερευνητές από διάφορους επιστημονικούς κλάδους επιχείρησαν να ενώσουν τα αντικείμενα του ενδιαφέροντός τους. Η συνεργασία αυτή βρήκε πρόσφορο έδαφος στο πεδίο της ΕΔ, βασιζόμενη στην εφαρμογή μεθοδολογιών και αλγορίθμων που είχαν ήδη χρησιμοποιηθεί από τους ερευνητές.

Πιο συγκεκριμένα η ΕΔ, χρησιμοποιεί έννοιες όπως δειγματοληψία, εκτίμηση και έλεγχος υποθέσεων από τη Στατιστική, καθώς και εφαρμογές όπως αναζήτηση αλγορίθμων, τεχνικές δημιουργίας υποδειγμάτων, θεωρίες τεχνητής νοημοσύνης, αναγνώρισης προτύπων και μηχανικής εκμάθησης.

Επιπλέον, υπάρχουν αρκετοί άλλοι τομείς των επιστημών που στήριξαν την πρόοδο της ΕΔ, όπως για παράδειγμα, η τεχνολογία των βάσεων δεδομένων. Τέλος, τεχνικές υψηλής απόδοσης από υπολογιστικής πλευράς και σχετικές με την ταξινόμηση παρέχουν βοήθεια σε σχέση με τη διαχείριση του μεγέθους και της συλλογής των τεράστιων συνόλων δεδομένων.

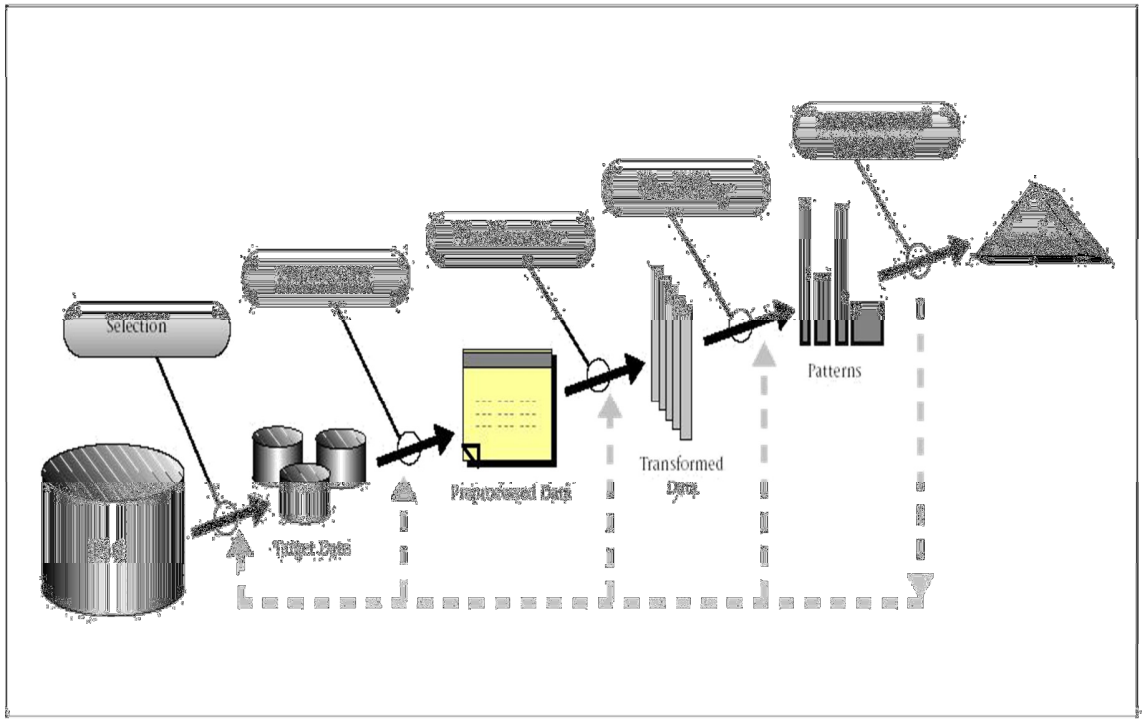
1.3.3 Η διαδικασία KDD

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων, είναι πιθανό να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις, αλληλεξάρτηση ή ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία μπορεί να μην είναι άμεσα εμφανή. Το είδος αυτό της γνώσης θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμο, αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Υπό αυτές τις συνθήκες, κρίνεται απαραίτητη η «μη επιβλεπόμενη» ανάκτηση γνώσης, που υποστηρίζεται από την εφαρμογή αλγορίθμων. Αυτήν την ανάγκη έρχεται να καλύψει η ΕΔ, η οποία αποτελεί τον πυρήνα της γενικότερης μεθοδολογίας της ανακάλυψης της γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases – KDD).

Η KDD είναι μία αυτοματοποιημένη διαδικασία, μέσω της οποίας γίνεται προσπάθεια διερευνητικής ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων, με στόχο τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά.
Ένας γενικός ορισμός της διαδικασίας KDD που ερμηνεύει με σαφήνεια τον όρο αυτό είναι:

«KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα».



Η ονομασία αυτή της KDD χρησιμοποιείται από το 1989 (πρώτο συνέδριο KDD) με στόχο να φανεί ότι η γνώση είναι το τελικό προϊόν μιας ανακάλυψης καθοδηγούμενης από τα δεδομένα. Με βάση τη σχετική βιβλιογραφία, θα διαχωρίσουμε τη διαδικασία KDD σε εννέα βήματα, τα οποία είναι:

- Την ανάπτυξη και κατανόηση της περιοχής της εφαρμογής, της σχετικά προγενέστερης γνώσης του προς εξέταση τομέα και τους στόχους του τελικού χρήστη.

- Την επιλογή και δημιουργία ενός κατάλληλου συνόλου δεδομένων. Την ολοκλήρωση δηλαδή των δεδομένων. Υπάρχουν διαφορετικά είδη αποθηκών πληροφοριών που μπορούν να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης. Κατά συνέπεια, οι πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο τελικά η διαδικασία εξόρυξης θα εκτελεστεί.

- Την δημιουργία στόχου – συνόλου δεδομένων. Επιλογή του συνόλου δεδομένων (δηλαδή μεταβλητές, δείγματα δεδομένων) στο οποίο η διαδικασία εξόρυξης πρόκειται να εκτελεσθεί.

- Τον καθαρισμό και την προ-επεξεργασία δεδομένων. Αυτό το βήμα περιλαμβάνει βασικές διαδικασίες όπως η αφαίρεση θορύβου ή των outliers, η συλλογή των απαραιτήτων πληροφοριών για την διαμόρφωση ή τη μέτρηση του θορύβου, η απόφαση σχετικά με τις στρατηγικές διαχείρισης των ελλειπόντων πεδίων δεδομένων.

- Τον μετασχηματισμό των δεδομένων. Τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Χρήση των μεθόδων μείωσης διαστάσεων ή μετασχηματισμού για τη μείωση του αριθμού των υπό εξέταση μεταβλητών ή την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.

- Την επιλογή των στόχων και των αλγορίθμων κατάλληλης μεθόδου εξόρυξης δεδομένων. Σε αυτό το βήμα αποφασίζουμε το στόχο της διαδικασίας KDD, επιλέγοντας τους στόχους εξόρυξης δεδομένων που θέλουμε να επιτύχουμε. Επίσης, επιλέγονται οι μέθοδοι που θα χρησιμοποιηθούν. Αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου και παραμέτρων. Επίσης η μέθοδος εξόρυξης δεδομένων πρέπει να αντιστοιχηθεί με τις απαιτήσεις και τα γενικά κριτήρια της διαδικασίας KDD.

- Την εξόρυξη δεδομένων. Εφαρμόζουμε ευφυείς μεθόδους, ψάχνουμε για ενδιαφέροντα πρότυπα γνώσης. Τα πρότυπα θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου τέτοιωναντιπροσωπεύσεων, όπως κανόνες κατηγοριοποίησης, δέντρα, παλινδρόμηση συσταδοποίηση κ.τ.λ. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.

- Την αξιολόγηση των προτύπων. Τα εξαγόμενα πρότυπα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα αληθινά ενδιαφέροντα πρότυπα.

- Την σταθεροποίηση και την παρουσίαση της γνώσης. Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα ή απλά την απεικόνισής μας και κάποιες τεχνικές αντιπροσώπευσης γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξορυγμένη γνώση στο χρήστη.

Η διαδικασία KDD θεωρείται διαλογική και επαναληπτική, δηλαδή μπορεί να απαιτηθεί η επιστροφή σε ένα προηγούμενο βήμα. Ως εφαρμογές της KDD στον χώρο των επιχειρήσεων αναφέρουμε τις δραστηριότητες σε marketing, επενδύσεις, προσδιορισμό απειλών, βιομηχανική παραγωγή, τηλεπικοινωνίες, καθαρισμό δεδομένων. Προφανώς, η δράση u964 της KDD σε αυτούς τους τομείς γίνεται μέσω της ΕΔ, δηλαδή η ΕΔ αποτελεί το εργαλείο της KDD.

Για να είναι σαφής η διαφορά μεταξύ διαδικασίας και εργαλείων, αναφέρουμε ότι ο όρος KDD χρησιμοποιείται για την περιγραφή ολόκληρης της διαδικασίας ανακάλυψης γνώσης από ένα σύνολο δεδομένων, ενώ ο όρος ΕΔ αναφέρεται στις τεχνικές που

χρησιμοποιούνται για την ανακάλυψη της γνώσης. Ο όρος ΕΔ αντιπροσωπεύει καλύτερα τη διαδικασία εύρεσης δομών γνώσης που περιγράφουν με ακρίβεια σύνολα πρωτογενών δεδομένων. Οι δομές αυτές αναδεικνύουν κρυμμένη γνώση (συνάφειες / κανόνες) που δεν είναι άμεσα ορατή και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων.

Κεφάλαιο 2ο

2.1 Market Basket Analysis

Η τεχνική Market Basket Analysis βασίζεται στην θεωρία ότι αν κάποιος πελάτης αγοράσει κάποιο συγκεκριμένο προϊόν (ή σύνολο προϊόντων), τότε είναι πολύ πιθανό (ή αντίστοιχα ελάχιστα πιθανό) να αγοράσει και ένα άλλο προϊόν (ή σύνολο προϊόντων). Το σύνολο των προϊόντων που αγοράζει ένας πελάτης κατά την διάρκεια μιας συγκεκριμένης αγοράς του ονομάζεται item set. Άρα, ένα item set αποτελείται από κάποια προϊόντα. Η τεχνική market basket analysis έχει σαν κύριο στόχο την ανάλυση των δεδομένων που προκύπτουν από τις αγορές των πελατών, με σκοπό την ανακάλυψη συσχετίσεων μεταξύ των διαφόρων προϊόντων. Τυπικά, μια συσχέτιση μεταξύ δύο προϊόντων είναι της μορφής:

- *IF { προϊόν A } THEN { προϊόν B }*

Η πιο πάνω συσχέτιση, δείχνει την σχέση μεταξύ των δύο προϊόντων. Φυσικά, μια συσχέτιση θα μπορούσε να περιλαμβάνει, αντί για μεμονωμένα προϊόντα, σύνολα προϊόντων. Μια τέτοια συσχέτιση θα μπορούσε να είναι η εξής:

- *IF { γάλα, ψωμί } THEN { βούτυρο, δημητριακά }*

Με τέτοιες συσχετίσεις συνδέονται άμεσα στατιστικές μεταβλητές, οι οποίες ονομάζονται support και confidence. Οι συσχετίσεις μεταξύ των προϊόντων ονομάζονται και κανόνες συσχέτισης, όπου οι μεταβλητές support και confidence δίνουν στατιστική πληροφορία που αφορά τους κανόνες αυτούς. Στην επόμενη παράγραφο δίνουμε έναν πιο τυπικό ορισμό της διαδικασίας market basket analysis.

Ο όρος Market Basket Analysis (MBA), ή ανάλυση καλαθιού αγορών, αφορά την ανάλυση διαφόρων υποσυνόλων αντικειμένων (προϊόντων), τα οποία επιλέχθηκαν μέσα από κάποιον μεγαλύτερο πληθυσμό αντικειμένων. Ένα παράδειγμα κανόνα είναι το A→B, όπου υποδηλώνει ότι: «εάν το αντικείμενο A υπάρχει στο καλάθι αγορών (market basket), τότε

υπάρχει και το αντικείμενο Β». Το Α ονομάζεται προηγηθέν αντικείμενο (antecedent item), ενώ το Β συνεπακόλουθο (consequent).

Σε έναν κανόνα MBA, όπως τον Α→Β, μπορεί να έχουμε ότι ενώ ο κανόνας είναι αληθείς (true), να ισχύει Α αληθές και Β μη αληθές, δηλαδή Α AND NOT Β. Έχοντας δηλαδή έναν κανόνα, μπορεί αυτός ενώ είναι αληθές, δηλαδή ενώ ισχύει, τα επιμέρους στοιχεία του να μην ισχύουν. Αυτό συμβαίνει, επειδή οι κανόνες στο market basket analysis θεωρούνται να έχουν κάποιους βαθμούς συνέπειας άμεσα συσχετισμένους με αυτούς, Τέτοιοι είναι οι confidence και support στατιστικές.

Το support κάποιου κανόνα Α→Β ορίζεται σαν:

- *Αν τα Α και Β ισχύουν μαζί για τουλάχιστον Χ% των καλαθιών αγορών, τότε το support του κανόνα είναι το Χ.*

Το confidence κάποιου κανόνα Α→Β ορίζεται σαν:

- *Από όλα τα καλάθια αγορών που περιέχουν το Α, αν τουλάχιστον Χ% περιέχουν επίσης το Β, τότε το confidence του κανόνα είναι Χ.*

Σαν καλάθι αγορών ονομάζουμε μια συναλλαγή (transaction) κάποιου πελάτη. Αν έχουμε για παράδειγμα μια υπεραγορά, τότε σαν καλάθι αγορών ονομάζουμε το σύνολο των προϊόντων που αγοράστηκαν από κάποιον πελάτη σε μια συγκεκριμένη συναλλαγή αυτού με το κατάστημα. Οι πιο πάνω κανόνες χρησιμοποιώντας τον όρο συναλλαγή αντί για καλάθι αγορών, θα γίνουν:

Το support κάποιου κανόνα Α→Β ορίζεται σαν:

- *Αν τα Α και Β ισχύουν μαζί για τουλάχιστον Χ% των συναλλαγών, τότε το support του κανόνα είναι το Χ.*

Το confidence κάποιου κανόνα A→B ορίζεται σαν:

- *Από όλες τις συναλλαγές που περιέχουν το A, αν τουλάχιστον X% περιέχουν επίσης το B, τότε το confidence του κανόνα είναι X.*

Για σκοπούς marketing, το confidence διαβεβαιώνει ότι ο κανόνας ισχύει, δηλαδή είναι αληθείς, μέχρι κάποιο συγκεκριμένο σημείο, ή αλλιώς με κάποια συγκεκριμένη πιθανότητα. Χρησιμοποιώντας το confidence, μπορεί κάποιος να διαβεβαιώσει ότι κάποιος κανόνας ισχύει αρκετά συχνά, ώστε να παίξει σημαντικό ρόλο κατά την λήψη αποφάσεων. Για παράδειγμα, αν σε κάποια υπεραγορά κάποιος κανόνας ισχύει αρκετά συχνά ή με αρκετά μεγάλη πιθανότητα, τότε λόγω του κανόνα αυτού μπορεί να αποφασιστεί διαρρύθμιση των προϊόντων που εμφανίζονται σε αυτόν.

Η χρήση του confidence από την άλλη, περιλαμβάνει και κάποιο ρίσκο. Αυτό συμβαίνει, επειδή όταν το συνεπακόλουθο αντικείμενο κάποιου κανόνα είναι δημοφιλές γενικότερα στις συναλλαγές, τότε το confidence του κανόνα μπορεί να είναι αρκετά μεγάλο, άσχετα με το αν τα δύο αντικείμενα (προηγηθέν και συνεπακόλουθο) δεν συσχετίζονται στην πραγματικότητα σε τόσο μεγάλο βαθμό. Βλέποντας το θέμα και διαισθητικά, από τον ορισμό του confidence, έχουμε ότι εάν ένας κανόνας A→B έχει confidence για παράδειγμα 95, τότε σημαίνει ότι από όλες τις συναλλαγές που περιέχουν το A, τουλάχιστον 95% περιέχουν επίσης το B. Στην περίπτωση όμως που το B είναι πολύ δημοφιλές προϊόν στις συναλλαγές γενικότερα, τότε μπορεί μεν να εμφανίζεται σε πολύ υψηλό ποσοστό (95%) στις ίδιες συναλλαγές με το A, όμως στην πραγματικότητα δεν συσχετίζεται τόσο με το συγκεκριμένο προϊόν, όσο θα συσχετιζόταν αν το B εμφανιζόταν κυρίως μόνο στις συναλλαγές που περιλαμβάνουν το A (δηλαδή αν το B δεν ήταν δημοφιλές). Το support παρέχει ένα μέτρο για το πόσο συχνά ένας κανόνας συμβαίνει (σε πόσες συναλλαγές είναι αυτός αληθείς), στο σύνολο όλων γενικότερα των συναλλαγών. Χρησιμοποιώντας το

support, ένας αναλυτής μπορεί να συμπεράνει κατά πόσο αξίζει την προσοχή του κάποιος κανόνας.

2.2. Κανόνες συσχέτισης - Association Rules

Οι κανόνες συσχέτισης, ή αλλιώς association rules, είναι κανόνες οι οποίοι εκφράζουν συσχετίσεις μεταξύ αντικειμένων. Πιο συχνά χρησιμοποιούνται σε συστήματα σημείων πώλησης προϊόντων, οπότε οι συσχετίσεις που εκφράζουν είναι μεταξύ των διαφόρων προϊόντων που αγοράζουν οι πελάτες. Οι κανόνες προκύπτουν με την διαδικασία εξόρυξης κανόνων συσχέτισης (association rule mining). Η εξόρυξη κανόνων συσχέτισης είναι μια πολύ διαδεδομένη διαδικασία, κατά την οποία γίνεται κατάλληλη ανάλυση των δεδομένων που είναι αποθηκευμένα σε βάσεις δεδομένων, για ανακάλυψη χρήσιμων συσχετίσεων μεταξύ προϊόντων, όπως για παράδειγμα η εύρεση προϊόντων τα οποία αγοράστηκαν μαζί.

Οι κανόνες συσχέτισης χρησιμοποιούνται σε εφαρμογές Market Basket Analysis. Όπως είδαμε σε προηγούμενη παράγραφο, με το όρο Market Basket Analysis εννοούμε την αναγνώριση διαφόρων ευκαιριών για πώληση προϊόντων που σχετίζονται μαζί (cross selling opportunities)., όπως για παράδειγμα:

- Αναγνώριση ομάδων προϊόντων που αγοράζονται μαζί (product baskets)
- Πρόβλεψη άλλων προϊόντων που ενδεχομένως να μπορούν να αγοραστούν μαζί, δεδομένων των προϊόντων που έχουν είδη αγοραστεί μαζί.

Δίνοντας πιο τυπικό ορισμό των κανόνων συσχέτισης, έχουμε:

A→B: *Δεδομένης της αγοράς του προϊόντος Α, υπάρχει μεγάλη πιθανότητα να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν Β. Το προϊόν Α ονομάζεται προηγηθέν, ενώ το Β συνεπακόλουθο.*

Για παράδειγμα, εάν ο κανόνας

- *{κρεμμύδια, λαχανικά} → {βοδινό}*

εξαγόταν από τα δεδομένα συναλλαγών πελατών σε μια υπεραγορά, αυτό θα σήμαινε ότι δεδομένης της αγοράς κρεμμυδιών και λαχανικών από κάποιον πελάτη, υπήρχε μεγάλη πιθανότητα να είχε αγοραστεί ή να υπήρχε μεγάλο ενδιαφέρον για βοδινό. Τέτοια πληροφορία μπορεί να χρησιμοποιηθεί σαν βάση για λήψη αποφάσεων που σχετίζονται με το marketing, όπως η τιμολόγηση και η τοποθέτηση των προϊόντων στους χώρους της υπεραγοράς.

«Το 90% των συναλλαγών που περιλαμβάνουν ψωμί και βούτυρο, περιλαμβάνουν και γάλα»

Στον πιο πάνω κανόνα, το ψωμί και το βούτυρο είναι τα προηγηθέντα προϊόντα, ενώ το γάλα το συνεπακόλουθο προϊόν. Επίσης, το confidence του κανόνα είναι 90 και εκφράζει την δύναμη του κανόνα. Στην περίπτωση αυτή, εκφράζει το ποσοστό των συναλλαγών που περιλαμβάνουν γάλα, δεδομένου ότι περιλαμβάνουν ψωμί και βούτυρο. Πολλοί άλλοι κανόνες συσχέτισης μπορούν να προκύψουν που να αφορούν τα ποιό πάνω προϊόντα, οι οποίοι θα προσφέρουν πολύτιμη πληροφορία για τα προϊόντα αυτά:

- *Να βρούμε όλους τους κανόνες που έχουν το γάλα σαν συνεπακόλουθο προϊόν.*

Αποτέλεσμα: Βλέπουμε με ποια προϊόντα συσχετίζεται το γάλα.

- *Να βρούμε όλους τους κανόνες με το ψωμί σαν προηγηθέν προϊόν.*

Αποτέλεσμα: Βλέπουμε ποια προϊόντα θα επηρεαστούν αν σταματήσει η πώληση του ψωμιού στην υπεραγορά. (Στο παράδειγμα το γάλα και το βούτυρο)

- *Να βρούμε όλους τους κανόνες με το γάλα σαν συνεπακόλουθο και το ψωμί σαν προηγηθέν.*

Αποτέλεσμα: Βλέπουμε ποια προϊόντα πρέπει ή είναι καλό να πωλούνται μαζί με το ψωμί (π.χ. βούτυρο), με σκοπό την μεγαλύτερη πιθανότητα για πώληση γάλακτος.

- *Να βρούμε όλους τους κανόνες που σχετίζονται με προϊόντα που βρίσκονται στα ράφια Α και Β στην υπεραγορά.*

Αποτέλεσμα: Shelf Planning Σχεδιάζουμε ποια προϊόντα είναι κατάλληλα για να τοποθετηθούν γειτονικά στα ράφια της υπεραγοράς. Προϊόντα τα οποία σχετίζονται στενά, δηλαδή με κανόνες συσχέτισης με μεγάλο παράγοντα confidence, θα ήταν καλή τακτική να τοποθετηθούν σε γειτονικά ράφια.

- *Τέλος χρήσιμο είναι να βρούμε τους καλύτερους k κανόνες, που έχουν κάποιο προϊόν (π.χ. το γάλα) σαν συνεπακόλουθο προϊόν. Ο όρος καλύτερος κανόνας αναφέρεται στον κανόνα με το μεγαλύτερο confidence factor ή την μεγαλύτερη συχνότητα εμφάνισης.*

Αποτέλεσμα: Κάνουμε χρήση μόνο αυτών των κανόνων οι οποίοι είναι και οι πλέον κατάλληλοι για λήψη αποφάσεων, αφού οι κανόνες αυτοί είναι οι πιο «βάσιμοι» και χαρακτηρίζουν το μεγαλύτερο μέρος των συναλλαγών (εφόσον είναι οι πιο συχνά εμφανιζόμενοι).

Από τα πιο πάνω, συμπεραίνουμε ότι οι κανόνες συσχέτισης μπορούν να προσφέρουν αξιοποιήσιμη πληροφορία που σχετίζεται με τα προϊόντα. Με χρήση της πληροφορίας αυτής, μπορεί να γίνει καλύτερη τιμολόγηση των προϊόντων, να αποφασιστούν ποια προϊόντα θα βγουν σε εκπτώσεις και προσφορές και ποια όχι, να μελετηθούν οι επιπτώσεις σε άλλα προϊόντα από τυχόν κατάργηση κάποιου προϊόντος, να αποφασιστεί η διαρρύθμιση των προϊόντων στα ράφια κ.α.

Επιπλέον, εκτός από το market basket analysis, οι κανόνες συσχέτισης χρησιμοποιούνται και σε άλλες εφαρμογές, όπως το Web usage mining, intrusion detection και βιοπληροφορική (bioinformatics).

2.3. Εξαγωγή κανόνων συσχέτισης

Ξεκινώντας ας δώσουμε τους ορισμούς που αφορούν στην Τεχνολογία της Επιστήμης, στην Εξόρυξη Δεδομένων καθώς και στην Εύρεση κανόνων συσχέτισης.

Σαν Επιστήμη της Τεχνολογίας καλείται η μελέτη των θεωρητικών θεμελίων της πληροφορίας και των υπολογιστών καθώς και των πρακτικών τεχνικών που απαιτούνται για την εκτέλεση και εφαρμογή τους στα υπολογιστικά συστήματα. Συνήθως περιγράφεται σαν μια συστηματική μελέτη των διαδικασιών που ακολουθούν οι αλγόριθμοι οι οποίοι δημιουργούν, περιγράφουν και μετατρέπουν και ανακατασκευάζουν την πληροφορία. Σαν Εξόρυξη Δεδομένων καλείται η διαδικασία της εξαγωγής κανόνων – προτύπων από τα δεδομένα. Η εξόρυξη Δεδομένων θεωρείται σαν ένα πολύ σημαντικό εργαλείο το οποίο χρησιμοποιείται από τις νεότερες επιχειρήσεις με σκοπό την μετατροπή των δεδομένων σε σημαντική πληροφορία.

Η Διαδικασία Εύρεσης κανόνων συσχέτισης – Discovery Association Rules in Data Mining, αποτελεί μια πολύ δημοφιλή και εμπλουτισμένη μέθοδος για την ανακάλυψη συσχετίσεων μεταξύ μεταβλητών σε μεγάλες βάσεις Δεδομένων. Οι Piatesky-Shapiro περιγράφουν την ανάλυση και την παρουσίαση των ισχυρών κανόνων οι οποίοι ανακαλύπτονται στις βάσεις δεδομένων χρησιμοποιώντας διαφορετικά μέτρα «ενδιαφέροντος». Βασιζόμενοι στην ιδέα των ισχυρών κανόνων ο Agrawal, παρουσίασε κανόνες συσχέτισης για την ανακάλυψη κανόνων μεταξύ προϊόντων σε μεγάλες τάξεις συναλλαγών οι οποίες καλούνται σαν point of sale συστήματα σε supermarket δηλαδή σε μεγάλους χώρους συναλλαγών.

Για παράδειγμα ο κανόνας {κρεμμύδια, πατάτες βοδινό κρέας} το οποίο προκύπτει συνήθως σε αγορές στο supermarket θα μπορούσε να υποδηλώσει ότι αν κάποιος πελάτης αγοράζει ταυτόχρονα κρεμμύδια και πατάτες είναι πολύ πιθανό να αγοράζει και βοδινό κρέας. Μια τέτοιου είδους πληροφορία μπορεί να χρησιμοποιηθεί σαν τη βάση- θεμέλιο για την απόφαση κάποιων ενεργειών προώθησης προϊόντων όπως για παράδειγμα αλλαγές στην τιμολόγηση ή τοποθέτηση σε κατάλληλη- ιδανική θέση. Επιπρόσθετα οι κανόνες συσχέτισης έχουν εφαρμογή και σε άλλους τομείς όπως εξόρυξη στο Διαδίκτυο και στην Βιοπληροφορική.

Ο Apriori στην πληροφορική και στην διαδικασία εξόρυξης δεδομένων αποτελεί έναν αλγόριθμο ο οποίος αφορά στην γνώση και εύρεση των κανόνων συσχέτισης.

Έχει σχεδιαστεί με τέτοιο τρόπο ώστε να μπορεί να εφαρμοστεί σε βάσεις δεδομένων οι οποίες περιέχουν συναλλαγές(όπως για παράδειγμα σύνολα προϊόντων που αγοράστηκαν από πελάτες ή λεπτομέρειες για την συχνή επίσκεψη σε έναν δικτυακό τόπο) Άλλοι αλγόριθμοι είναι σχεδιασμένοι για να εφαρμόζονται, για την εύρεση κανόνων συσχέτισης, σε δεδομένα στα οποία δεν εμπλέκεται κανενός είδος συναλλαγή ή σε δεδομένα στα οποία δεν υπάρχει συγκεκριμένο χρονοδιάγραμμα για παράδειγμα στην DNA αλληλουχία.

Όπως είναι σύνηθες στην εξόρυξη κανόνων συσχέτισης, δίνεται ένα δεδομένο σύνολο από itemesets, (για παράδειγμα σύνολα από πωλήσεις λιανικής όπου σε καθεμία δημιουργείται μια μεμονωμένη λίστα με τα προϊόντα τα οποία αγοράστηκαν) και σε αυτό το σύνολο ο αλγόριθμος επιχειρεί να βρει όλα τα υποσύνολα τα οποία είναι κοινά σε τουλάχιστον ένα ελάχιστο αριθμό c των itemsets. Ο Apriori αλγόριθμος χρησιμοποιεί τις τεχνικές breadth-first search και την tree τεχνική με στόχο να υπολογίσει τα υποψήφια Item sets σε ικανοποιητικό βαθμό. Παράγει υποψήφια item sets με μέγεθος k από Item sets με μέγεθος k-1.

Στην συνέχεια «κόβει» τα υποψήφια item sets τα οποία δεν έχουν συχνή εμφάνιση.

2.3.1 Συσχετίσεις – MBA

Η πιο διαδεδομένη μέθοδος για παραγωγή αλληλοσυσχετίσεων και αλληλεξαρτήσεων αποτελεί η εύρεση κανόνων συσχέτισης. Το πρόβλημα εξαγωγής κανόνων συσχέτισης παρουσιάστηκε αρχικά το 1993 ως μια προσπάθεια εξαγωγής χρήσιμων συσχετισμών μεταξύ των πεδίων μιας βάσης δεδομένων.

Όπως αναφέραμε σε προηγούμενο κεφάλαιο η πιο συνηθισμένη εφαρμογή της μεθόδου της συσχέτισης είναι η «Ανάλυση του καλαθιού της νοικοκυράς». (market basket analysis). Σκοπός είναι αν αναγνωρισθούν τα αγαθά τα οποία αγοράστηκαν μαζί. Συγκεκριμένα ένας κανόνας συσχέτισης θα μπορούσε να πει ότι το γάλα πωλείται μαζί με το τυρί, με την προφανή αξιοποίηση της πληροφορίας που είναι η τοποθέτηση και των δυο αυτών προϊόντων στο ίδιο σημείο πώλησης.

Μερικές συνήθεις πρακτικές εφαρμογές τους είναι η εύρεση προϊόντων που πωλούνται μαζί σε μια συναλλαγή, η επεξεργασία ερωτηματολογίων, η εύρεση των προϊόντων που διακινούνται μαζί σε μια αποθήκη για πρόβλεψη προμήθειας καθώς και η εύρεση των λέξεων που συναντώνται μαζί σε ένα κείμενο.

¨Όπως βέβαια είπαμε και προηγούμενα, ένας κανόνας συσχέτισης είναι μια έκφραση της μορφής X Y , όπου X και Y είναι σύνολα τιμών των πεδίων, όπως για παράδειγμα σύνολα προϊόντων. Η σπουδαιότητα ενός κανόνα συσχέτισης καθορίζεται αναλογικά από το ποσοστό εφαρμογής του κανόνα επί του συνόλου εκπαίδευσης.

Συγκεκριμένα οι αλγόριθμοι συσχέτισης που έχουν προταθεί και εφαρμόζονται πρακτικά, εξάγουν κανόνες συσχέτισης της μορφής: «Το 98% των πελατών που αγοράζουν γάλα και κρέας αγοράζουν επίσης και τυρί. Αλλά στο 70% των αγορών έχουν αγορασθεί γάλα και τυρί και κρέας». Το πρώτο ποσοστό αναφέρεται ως αξιοπιστία (confidence) του κανόνα ενώ το δεύτερο ως επιβεβαίωση (support).

Η επιβεβαίωση αφορά στο ποσοστό που εμφανίζονται και τρία αγαθά μαζί επί του συνόλου εκπαίδευσης ενώ η αξιοπιστία αφορά στο ποσοστό που εμφανίζονται τα αγαθά επί του αριθμού αγορών που περιέχουν γάλα και κρέας. Το πρόβλημα εύρεσης κανόνων συσχέτισης εστιάζεται στην εύρεση όλων των κανόνων που έχουν μια καθορισμένη από τον χρήστη ελάχιστη τιμή επιβεβαίωσης και αξιοπιστίας.

Κεφάλαιο 3ο

3.1 Apriori αλγόριθμος

Ηπαρουσίαση του αλγορίθμου Apriori, οι συμβολισμοί, οι ορισμοί καθώς και τα παραδείγματα που χρησιμοποιούμε προέρχονται από την πηγή.

Ο αλγόριθμος Apriori δέχεται ως είσοδο ένα σύνολο αγορών (transactions) που αποτελεί και το σύνολο εκπαίδευσης. Κάθε αγορά είναι ουσιαστικά μια λίστα (itemset) από αγαθά (items) τα οποία αγοράστηκαν μαζί.

3.2 Ορισμοί

Itemset: Σύνολο από αντικείμενα – αγαθά (items).

K-itemset: Σύνολο από k αντικείμενα- αγαθά

Σύνολο από k-itemsets: Σύνολο από k -άδες items. Το σύνολο αυτό περιέχει υποσύνολα, όπου κάθε υποσύνολο περιέχει k - items το καθένα.π.χ. Ένα 3-itemset: {I1, I2, I3}

Σύνολο από 3-itemsets: { {I1, I2, I3} {I1, I2, I5} {I2, I4, I5} }

Minimum support threshold (min_sup): Κατώτερο όριο το οποίο πρέπει να ικανοποιούν τα itemsets για να είναι frequent. Ένα itemset ικανοποιεί το κατώφλι και είναι frequent, αν ο αριθμός εμφανίσεων του στην Βάση Δεδομένων είναι μεγαλύτερος ή ίσος από το κατώτερο όριο αυτό.

Frequent itemset: Ένα itemset I είναι frequent, αν ο αριθμός εμφανίσεων του στην Βάση Δεδομένων είναι μεγαλύτερος ή ίσος από το κατώτερο όριο του minimum support. (Αυτό δηλώνεται από το L_i για το i-στο itemset). $P(I) \geq min\_sup$

Σύνολο Lk: σύνολο από frequent k-itemsets

Σύνολο Ck: σύνολο από υποψήφια frequent k-itemsets

Apriori Property: Κάθε υποσύνολο των frequent itemset πρέπει να είναι frequent.

Join Operation- λειτουργία Συνένωσης: Εύρεση του L_k, του συνόλου δηλαδή των παραγώγων k-itemsets το οποίο προκύπτει από την συνένωση των L_{k-1} μεταξύ τους.

3.3 Γενική εισαγωγή στον Apriori

Πρόκειται για τον βασικό αλγόριθμο για εύρεση frequent itemsets. Τα frequent itemsets μας είναι χρήσιμα, επειδή από αυτά προκύπτουν με κατάλληλες μεθόδους οι κανόνες συσχέτισης. Επίσης, υπάρχουν και αλγόριθμοι οι οποίοι αποτελούν βελτιώσεις του Apriori. Στα πλαίσια της διπλωματικής αυτής, οι αλγόριθμοι αυτοί δεν μελετήθηκαν, αφού ο Apriori αλγόριθμος μπορεί να μας δώσει τα επιθυμητά αποτελέσματα. Το πρόβλημα της εύρεσης των κανόνων συσχέτισης που έχουν την επιθυμητή επιβεβαίωση και αξιοπιστία μπορεί να διαιρεθεί σε δυο υπο-προβλήματα:

Εύρεση όλων των συνδυασμών των προϊόντων που έχουν επιβεβαίωση πάνω από την ελάχιστη επιβεβαίωση (minimum support). Όλοι αυτοί οι συνδυασμοί ονομάζονται μεγάλες λίστες από προϊόντα (large itemsets) και όλοι οι υπόλοιποι συνδυασμοί μικρές λίστες από προϊόντα(small itemsets).

Χρήση όλων των μεγάλων λιστών από προϊόντα για εξόρυξη των κανόνων συσχέτισης που ικανοποιούν την ελάχιστη αξιοπιστία. Για παράδειγμα, έστω τα ABED και AB είναι μεγάλες λίστες από προϊόντα. Μπορούμε να καθορίσουμε αν ο κανόνας συσχέτισης AB CD ξεπερνά την ελάχιστη αξιοπιστία, υπολογίζοντας το λόγο:

r= επιβεβαίωση (ABCD)/ επιβεβαίωση (AB)

Αν r ελάχιστη αξιοπιστία, τότε ο κανόνας συσχέτισης γίνεται αποδεκτός.

Η εύρεση των μεγάλων λιστών από προϊόντα, για να αποφύγει κανείς ένα εξαντλητικό ψάξιμο όλων των δυνατών συνδυασμών, βασίζεται στο ότι: μια λίστα από προϊόντα είναι μεγάλη λίστα από προϊόντα αν κάθε υποσύνολό της είναι μεγάλη λίστα από προϊόντα. Apriori προτάθηκε από τους R .Agrawal και R .Srikant το 1994. Στόχος τους ήταν η εξόρυξη frequent itemsets για Boolean κανόνες συσχέτισης. Το όνομα βασίζεται στο γεγονός ότι ο

αλγόριθμος χρησιμοποιεί «προηγούμενη γνώση» (prior knowledge) ιδιοτήτων των frequent itemsets. Επίσης, ο αλγόριθμος χρησιμοποιεί την προσέγγιση level - wise search, κατά την οποία k-itemsets χρησιμοποιούνται για την εύρεση k+1-itemsets.

Ο Apriori είναι ένας πολύ σημαντικός αλγόριθμος ο οποίος χρησιμοποιείται στην εξόρυξη των απλούστερων μορφών από frequent patterns – itemsets, με στόχο την εξαγωγή κανόνων συσχέτισης.

Ξεκινώντας λοιπόν καλό θα ήταν να κάνουμε μια πολύ σύντομη παρουσίαση στον τρόπο που λειτουργεί ο αλγόριθμος.

- *Βρίσκουμε τα αγαθά που εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση (minimum support), δηλαδή το σύνολο L_1=μεγάλες λίστες από 1 – αγαθό (large 1 item sets)*
- *Από κ=2 και όσο L_{k-1} δεν είναι κενό κάνε:*

Βρες το σύνολο C_k όλων των υποψήφιων μεγάλων λιστών από κ-αγαθά (candidate large k-itemsets) με βάση το L_{k-1} .Βρες ποια από αυτά εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση και φτιάξε το σύνολο L_k = μεγάλες λίστες από κ-αγαθά.

- *Για κάθε στοιχείο των L_1, L_2, L_3,…… L_n βρες ποια ικανοποιούν την ελάχιστη αξιοπιστία (minimum confidence).*
- *Στην συνέχεια χρησιμοποιούμε αυτά τα frequent itemset για την παραγωγή κανόνων συσχέτισης.*

Βρίσκουμε το σύνολο των frequent 1-itemsets (k=1) που ικανοποιούν τον περιορισμό του min support, με ψάξιμο (scanning) όλης της βάσης δεδομένων και μέτρηση των εμφανίσεων τους. Το αποτέλεσμα είναι το σύνολο L_1.

Βρίσκουμε το σύνολο των frequent 2-itemsets (k=2) χρησιμοποιώντας το σύνολο L_2. Το αποτέλεσμα το ονομάζουμε σύνολο L_2.

Ακολουθώντας την ίδια διαδικασία, βρίσκουμε τελικά το σύνολο $L_κ$ χρησιμοποιώντας το σύνολο $L_{κ-1}$.

Για την εύρεση κάθε L_n με n=1…k, χρειάζεται ένα πλήρες ψάξιμο (full scanning) όλης της Βάσης Δεδομένων.

Βήματα αλγορίθμου Apriori

Ο αλγόριθμος βασίζεται σε δύο βήματα: το join step (βήμα συνένωσης) και το pruning step (βήμα κλαδέματος). Το σύνολο $L_κ$ προκύπτει μετά από τις διαδικασίες join και pruning δύο συνόλων $L_{κ-1}$ και $L_{κ-1}$.

Βήμα 1: join step

Όπως είπαμε και προηγουμένως, το σύνολο $L_κ$ θα βρεθεί από το σύνολο $L_{κ-1}$. Για να βρεθεί το $L_κ$, πρέπει πρώτα να βρεθεί το σύνολο C_k των υποψήφιων itemsets. Το σύνολο C_k θα βρεθεί αφού γίνει joined το $L_κ$ με τον εαυτό του.

- *$L_{κ-1}$. join $L_{κ-1}$.: Γίνεται μόνο στα itemsets του $L_{κ-1}$. που είναι joinable.*

- *Joinable: Δύο k-1-itemsets είναι joinable, εάν τα items τους μέχρι και το k-2 είναι τα ίδια.*

- *Το k-itemset αυτό αποτελεί υποψήφιο k-itemset για το σύνολο $L_κ$ και θα μπει στο σύνολο $C_κ$ των υποψήφιων k-itemsets.*

Παράδειγμα για k=4: Για να βρούμε το L_4 από το L_3 θα κάνουμε την πράξη L_3 join L_3: για κάθε itemset στο L_3 κοιτάμε αν είναι joinable με κάθε άλλο itemset στο L_3 (εκτός τον εαυτό του), δηλαδή αν έχουν τα ίδια items στις πρώτες 2 θέσεις. Αν δύο itemsets είναι joinable, τότε γίνονται joined και το 4-itemset που προκύπτει εισάγεται στο σύνολο C_4, αφού αποτελεί υποψήφιο 4-itemset για το L_4.

Βήμα 2: pruning step

Το σύνολο $C_κ$ από k-itemsets είναι υπερσύνολο του $L_κ$, το οποίο θέλουμε να βρούμε. Ισχύει ότι: τα k-itemsets-μέλη του $C_κ$ μπορεί να είναι η να μην είναι frequent, αλλά όλα τα frequent k-itemsets συμπεριλαμβάνονται στο $C_κ$. Το ποια από τα itemsets αυτά είναι frequent και ποια όχι, καθορίζεται με ένα ψάξιμο στην Βάση Δεδομένων για κάθε itemset και μέτρηση του πλήθους των εμφανίσεων του κάθε ενός από αυτά. Frequent είναι μόνο τα itemset που έχουν αριθμό εμφανίσεων μεγαλύτερο ή ίσο του minimum support.

Επειδή το C μπορεί να είναι τεράστιο, γίνεται μια διαδικασία ελαχιστοποίησης των itemsets του. Χρησιμοποιείται η Apriori ιδιότητα:

«Οποιοδήποτε k-1-itemset δεν είναι frequent, δεν μπορεί να είναι υποσύνολο κάποιου frequent k-itemset»

Έτσι, αν οποιοδήποτε από τα k-1-itemsets κάποιου υποψήφιου k-itemset στο $C_κ$ δεν υπάρχει στο $L_{κ-1}$ τότε το υποψήφιο k-itemset δεν μπορεί να είναι frequent και αφαιρείται από το $C_κ$. Με τον τρόπο αυτό, για κάθε τέτοιο k-itemset (που σίγουρα δεν είναι frequent), γλιτώνουμε την προσπέλαση στον πίνακα συναλλαγών και την μέτρηση του αριθμού των εμφανίσεων του, δηλαδή τις πράξεις που θα κάναμε για να δούμε αν θα ικανοποιούσε το minimum support.

Να τονίσουμε ότι με την pruning διαδικασία δεν προκύπτουν τα frequent k-itemsets, αλλά τα k-itemsets που αποκλείεται να είναι frequent. Έτσι, μετά το pruning, χρειάζεται η προσπέλαση της βάσης δεδομένων για κάθε ένα από τα εναπομείναντα itemsets στο Ck, ώστε να καθοριστεί αν ικανοποιούν το minimum support. Μόνο αν το ικανοποιούν θα είναι frequent itemsets.

Κεφάλαιο 4ο

Περιγραφή δεδομένων

Τα δεδομένα που συγκεντρώθηκαν με στόχο την επεξεργασία και ανάλυσή τους αφορούν πελάτες που κανουν τις αγορές τους σε mini market τους μήνες Σεμπτέμβρη έως και τον Νοέμβριο του 2017. Στην αρχική τους μορφή τα δεδομένα συγκεντρώθηκαν σε αρχείο δεδομένων τύπου excel και περιλάμβαναν τα παρακάτω πεδία.

Ώρες	
a	8-12
b	12-15
c	17-20
Ηλικία	
o (old)	50-70
m (middle)	30-50
y (young)	18-30
Προϊόντα	
x1	Χαρτικα
x2	Απορρυπαντικά
x3	Καλλυντικά
x4	Μαλακτικά
x5	Μιας χρησης
x6	Περιποιησης
x7	Καθαρισμού
x8	Σπιτιού
x9	Βρεφικα
x10	Διακοσμισης

Προκειμένου να εξασφαλίσουμε ένα ποιοτικό σύνολο δεδομένων απαλλαγμένο από προβλήματα όπως για παράδειγμα ελλειπή δεδομένα (μη συμπληρωμένα ή διαγραμμένα πεδία), λανθασμένα δεδομένα (λανθασμένες τιμές ή ακραίες τιμές), ασυνέπειες δεδομένων (έλλειψη αναγνωριστικών ή κωδικοποίησης) προχωρήσαμε σε μια σειρά από ενέργειες που περιλαμβάνουν καθαρισμό, ολοκλήρωση, μείωση, μετασχηματισμό και διακριτοποίηση των δεδομένων .

Ένα απαραίτητο κομμάτι της ανάλυσης του συνόλου δεδομένων αποτελεί και ο καθαρισμός των δεδομένων, κατά τον οποίο εντοπίζονται και διορθώνονται ή αφαιρούνται οι ανακριβείς και κατεστραμμένες τιμές στο σύνολο των δεδομένων. Στην συγκεκριμένη περίπτωση όπου οι

εγγραφές γίνονταν με μη αυτοματοποιημένο τρόπο και χωρίς συγκεκριμένη μορφολογία και τυποποίηση εμφανίστηκαν πολλά κενά πεδία και λανθασμένες καταχωρίσεις. Όλα τα παραπάνω ως γνωστόν δύναται να επηρεάσουν αρνητικά την πορεία της ανάλυσης και να οδηγήσουν σε ψευδείς κανόνες κατά την διαδικασία της κατηγοριοποίησης.

Από το αρχικό σύνολο δεδομένων αφαιρέθηκαν τα πεδία που δεν μας βοηθούσαν στην ολοκήτωση του συνόλου. Στην συνέχεια αφαιρέθηκαν τα πεδία που εμφάνιζαν υψηλό ποσοστό μη συμπληρωμένων τιμών (της τάξης 70% περίπου), καθώς προκειμένου να εξαχθούν χρήσιμοι κανόνες θα πρέπει να εξασφαλίσουμε κάποια ελάχιστη υποστήριξη στο υπάρχον σύνολο δεδομένων.Δηλαδή στην περίπτωση που εξαχθεί κάποιος κανόνας να έχουμε ικανοποιητικό αριθμό περιπτώσεων στις οποίες να εφαρμόζεται αυτός ο κανόνας. Ο παραπάνω μετασχηματισμός των πεδίων κειμένου ήταν αναγκαίος διότι διαφορετικά , αφενός δεν υπήρχε συγκεκριμένο όφελος από τα δεδομένα των πεδίων και αφετέρου δεν ήταν δυνατή η μετατροπή του αρχείου δεδομένων τύπου excel σε arff (δεν δέχεται ελληνικούς χαρακτήρες και σύμβολα).

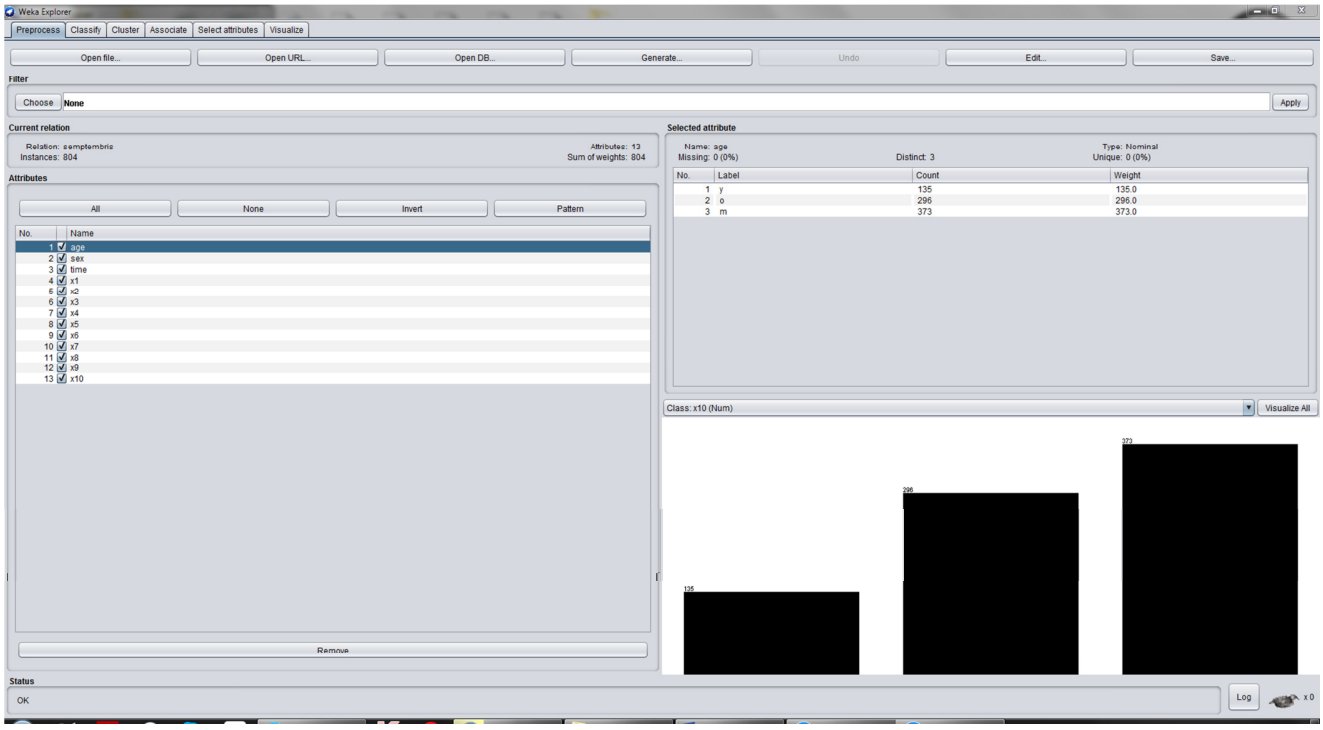
4.2 Δημιουργώντας το αρχείο δεδομένων

4.2.1ARFF

Στο τελικό σύνολο δεδομένων έγινε μετατροπή από την μορφή excel σε μορφή csv (comma separated value) δηλαδή σε αρχείο τύπου κειμένου με συγκεκριμένη μορφοποίηση όπου οι τιμές των μεταβλητών διαχωρίζονται με κόμμα. Κατόπιν χρησιμοποιώντας ένα πρόγραμμα επεξεργασίας κειμένου και με κατάλληλους μετασχηματισμούς δημιουργήθηκε στην κατάλληλη μορφοποίηση και τυποποίηση , όπως απεικονίζεται παρακάτω, το αρχείο δεδομένων σε μορφή arff , η οποία είναι και η επιθυμητή εισόδου για την επεξεργασία στο WEKA.

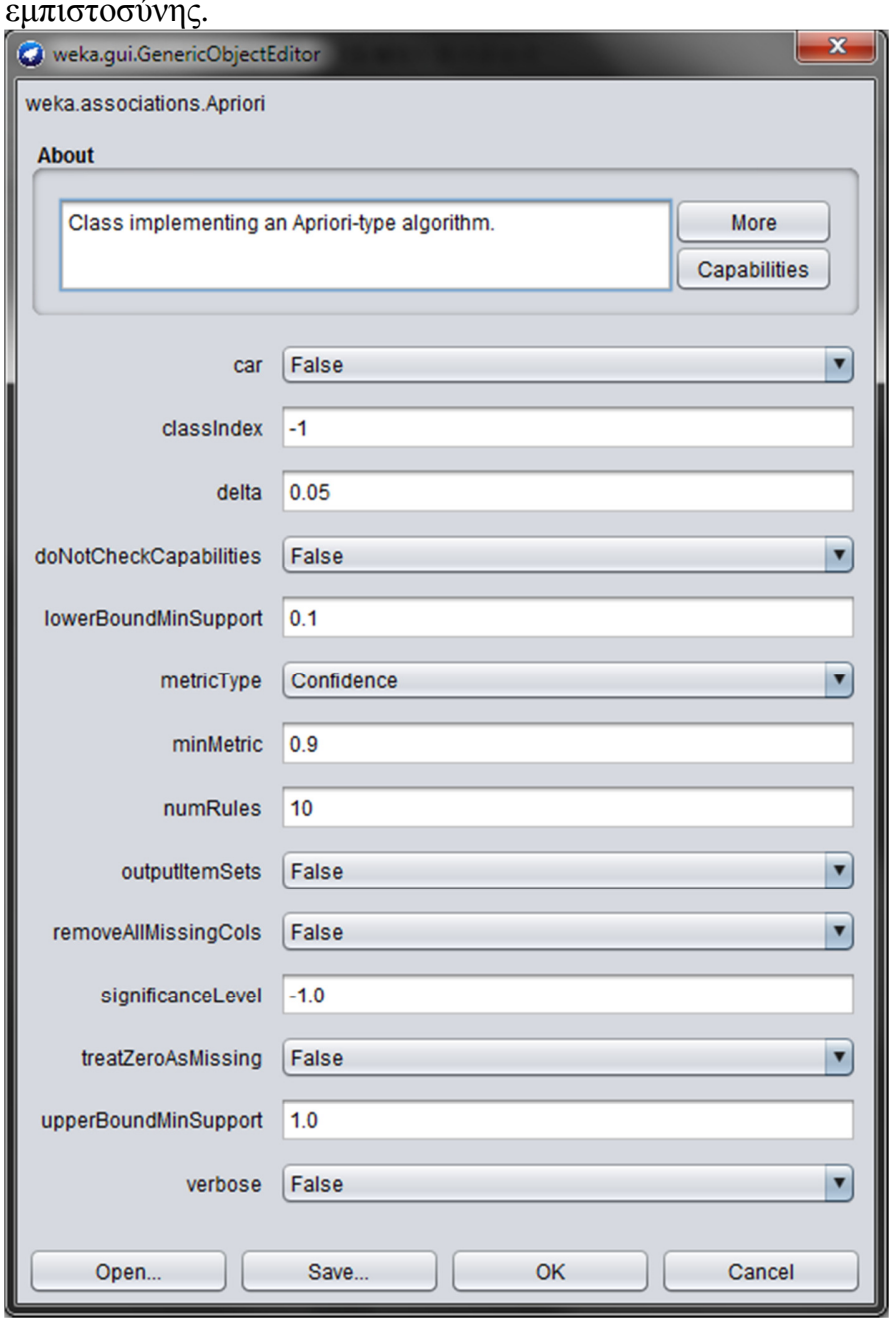
Αφού φορτωθεί το αρχείο δεδομένων στο WEKA, στην αρχική οθόνη του προγράμματος παρουσιάζονται πολλές χρήσιμες πληροφορίες για το σύνολο δεδομένων, όπως για παράδειγμα ο αριθμός εγγραφών, οι μεταβλητές και οι τύποι τους, το ποσοστό ελλειπουσών

τιμών και οπτικοποιείται σε γραφήματα ράβδων η ανάλυση τιμών ανά μεταβλητή.



4.3 Συναρτήσεις Apriori

Το WEKA περιλαμβάνει και αλγορίθμους για την εξόρυξη Κανόνων Συσχέτισης. Ο χρήστης μπορεί να βρει τους σχετικούς αλγορίθμους στο tab ″Associate″. Περιλαμβάνονται ορισμένοι αλγόριθμοι, μεταξύ των οποίων και ο Apriori. Τα δεδομένα πρέπει να είναι διακριτά. Με την εφαρμογή του Apriori μπορούν να βρεθούν κανόνες, οι οποίοι υπερβαίνουν τις ελάχιστες τιμές υποστήριξης και εμπιστοσύνης.



Στην παραπάνω εικόνα παρουσιάζεται το παράθυρο ρύθμισης παραμέτρων του Apriori. Στην υλοποίηση του Apriori, η οποία υπάρχει στο WEKA, εκτελείται μια επαναλαμβανόμενη διαδικασία, όπου ο αλγόριθμος μειώνει σταδιακά την τιμή της υποστήριξης, μέχρι να βρεί έναν προκαθορισμένο αριθμό κανόνων. Στο πεδίο '' upperBoundMinSupport '' ορίζεται η τιμή εκκίνησης για την ελάχιστη υποστήριξη, στο πεδίο ''lowerBoundMinSupport'' ορίζεται η μικρότερη δυνατή τιμή για την ελάχιστη υποστήριξη και στο πεδίο ''delta'' ορίζεται το βήμα μείωσης. Το πλήθος των κανόνων που θα εξαχθούν ορίζεται στο πεδίο ''numRules'' . Προβλέπονται διάφορες μετρικές (πεδίο '' metricType'') για την εξαγωγή των κανόνων, με προτεινόμενη επιλογή την εμπιστοσύνη. Στο πεδίο ''minMetric'' ορίζεται η ελάχιστη τιμή της μετρικής

4.3.1 Μέθοδος Confidence

Ο κανόνας του Confidence ορίζεται ως:

$$conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$$

Κατα την εκτέλεση του αλγοριθμου Confidence εμφανίζεται το αποτέλεσμα για δείγμα 100 πελατών.

```
=== Run information ===


Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     semptembris-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    804
Attributes:   13
              age
              sex
              time
              x1
              x2
              x3
```

```
        x4

        x5

        x6

        x7

        x8

        x9

        x10

=== Associator model (full training set) ===



Apriori

=======


Minimum support: 0.5 (402 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 10


Generated sets of large itemsets:


Size of set of large itemsets L(1): 11


Size of set of large itemsets L(2): 39


Size of set of large itemsets L(3): 52
```

Size of set of large itemsets L(4): 32

Size of set of large itemsets L(5): 3

Best rules found:

1. x2=0 708 ==> x4=0 684 <conf:(0.97)> lift:(1.06) lev:(0.05) [39] conv:(2.54)
2. x2=0 x6=0 496 ==> x4=0 478 <conf:(0.96)> lift:(1.06) lev:(0.03) [26] conv:(2.34)
3. x2=0 x10=0 657 ==> x4=0 633 <conf:(0.96)> lift:(1.06) lev:(0.04) [34] conv:(2.35)
4. x2=0 x7=0 431 ==> x4=0 415 <conf:(0.96)> lift:(1.06) lev:(0.03) [22] conv:(2.27)
5. x2=0 x9=0 613 ==> x4=0 590 <conf:(0.96)> lift:(1.06) lev:(0.04) [31] conv:(2.29)
6. x2=0 x3=0 x6=0 445 ==> x4=0 428 <conf:(0.96)> lift:(1.06) lev:(0.03) [22] conv:(2.21)
7. x2=0 x3=0 596 ==> x4=0 573 <conf:(0.96)> lift:(1.06) lev:(0.04) [30] conv:(2.22)
8. x2=0 x8=0 602 ==> x4=0 578 <conf:(0.96)> lift:(1.05) lev:(0.04) [29] conv:(2.16)
9. x2=0 x6=0 x10=0 450 ==> x4=0 432 <conf:(0.96)> lift:(1.05) lev:(0.03) [22] conv:(2.12)
10. x2=0 x9=0 x10=0 568 ==> x4=0 545 <conf:(0.96)> lift:(1.05) lev:(0.03) [27] conv:(2.12)
11. x2=0 x5=0 517 ==> x4=0 496 <conf:(0.96)> lift:(1.05) lev:(0.03) [25] conv:(2.1)
12. x2=0 x3=0 x10=0 547 ==> x4=0 524 <conf:(0.96)> lift:(1.05) lev:(0.03) [25] conv:(2.04)
13. x2=0 x6=0 x9=0 427 ==> x4=0 409 <conf:(0.96)> lift:(1.05) lev:(0.03) [20] conv:(2.01)
14. x2=0 x8=0 x10=0 565 ==> x4=0 541 <conf:(0.96)> lift:(1.05) lev:(0.03) [26] conv:(2.02)
15. x2=0 x3=0 x9=0 508 ==> x4=0 486 <conf:(0.96)> lift:(1.05) lev:(0.03) [23] conv:(1.98)
16. x2=0 x5=0 x10=0 479 ==> x4=0 458 <conf:(0.96)> lift:(1.05) lev:(0.03) [21] conv:(1.95)
17. x2=0 x8=0 x9=0 516 ==> x4=0 493 <conf:(0.96)> lift:(1.05) lev:(0.03) [23] conv:(1.93)
18. x2=0 x5=0 x9=0 448 ==> x4=0 428 <conf:(0.96)> lift:(1.05) lev:(0.03) [20] conv:(1.91)
19. x2=0 x3=0 x5=0 437 ==> x4=0 417 <conf:(0.95)> lift:(1.05) lev:(0.02) [19] conv:(1.86)

20. x2=0 x3=0 x8=0 492 ==> x4=0 469 <conf:(0.95)> lift:(1.05) lev:(0.03) [21] conv:(1.84)

21. x2=0 x3=0 x9=0 x10=0 465 ==> x4=0 443 <conf:(0.95)> lift:(1.05) lev:(0.02) [19] conv:(1.81)

22. x2=0 x8=0 x9=0 x10=0 485 ==> x4=0 462 <conf:(0.95)> lift:(1.05) lev:(0.03) [20] conv:(1.81)

23. x2=0 x5=0 x8=0 440 ==> x4=0 419 <conf:(0.95)> lift:(1.05) lev:(0.02) [18] conv:(1.79)

24. x2=0 x3=0 x8=0 x10=0 457 ==> x4=0 434 <conf:(0.95)> lift:(1.04) lev:(0.02) [17] conv:(1.71)

25. x8=0 x9=0 604 ==> x10=0 570 <conf:(0.94)> lift:(1.01) lev:(0.01) [6] conv:(1.16)

26. x5=0 x8=0 x9=0 455 ==> x10=0 429 <conf:(0.94)> lift:(1.01) lev:(0.01) [4] conv:(1.13)

27. x5=0 x8=0 523 ==> x10=0 493 <conf:(0.94)> lift:(1.01) lev:(0.01) [5] conv:(1.13)

28. x8=0 695 ==> x10=0 655 <conf:(0.94)> lift:(1.01) lev:(0.01) [6] conv:(1.14)

29. x2=0 x8=0 x9=0 516 ==> x10=0 485 <conf:(0.94)> lift:(1.01) lev:(0) [3] conv:(1.08)

30. x2=0 x5=0 x8=0 440 ==> x10=0 413 <conf:(0.94)> lift:(1.01) lev:(0) [2] conv:(1.06)

31. x2=0 x8=0 602 ==> x10=0 565 <conf:(0.94)> lift:(1.01) lev:(0) [3] conv:(1.06)

32. x2=0 x4=0 x8=0 x9=0 493 ==> x10=0 462 <conf:(0.94)> lift:(1) lev:(0) [2] conv:(1.03)

33. x4=0 x8=0 x9=0 537 ==> x10=0 503 <conf:(0.94)> lift:(1) lev:(0) [2] conv:(1.03)

34. x4=0 x8=0 625 ==> x10=0 585 <conf:(0.94)> lift:(1) lev:(0) [1] conv:(1.02)

35. x2=0 x4=0 x8=0 578 ==> x10=0 541 <conf:(0.94)> lift:(1) lev:(0) [1] conv:(1.02)

36. sex=f 435 ==> x10=0 407 <conf:(0.94)> lift:(1) lev:(0) [1] conv:(1.01)

37. x3=0 x8=0 x9=0 495 ==> x10=0 463 <conf:(0.94)> lift:(1) lev:(0) [1] conv:(1.01)

38. x4=0 x5=0 x8=0 458 ==> x10=0 428 <conf:(0.93)> lift:(1) lev:(0) [0] conv:(0.99)

39. x4=0 732 ==> x2=0 684 <conf:(0.93)> lift:(1.06) lev:(0.05) [39] conv:(1.78)

40. x3=0 x8=0 578 ==> x10=0 540 <conf:(0.93)> lift:(1) lev:(0) [0] conv:(1)

41. x3=0 x5=0 x8=0 437 ==> x10=0 408 <conf:(0.93)> lift:(1) lev:(0) [0] conv:(0.98)

42. x4=0 x10=0 678 ==> x2=0 633 <conf:(0.93)> lift:(1.06) lev:(0.04) [35] conv:(1.76)

43. x3=0 x4=0 x6=0 459 ==> x2=0 428 <conf:(0.93)> lift:(1.06) lev:(0.03) [23] conv:(1.71)

44. x5=0 603 ==> x10=0 562 <conf:(0.93)> lift:(1) lev:(-0) [0] conv:(0.96)

45. x9=0 704 ==> x10=0 656 <conf:(0.93)> lift:(1) lev:(-0) [0] conv:(0.96)

46. x7=0 x8=0 440 ==> x10=0 410 <conf:(0.93)> lift:(1) lev:(-0) [0] conv:(0.95)

47. x3=0 x4=0 616 ==> x2=0 573 <conf:(0.93)> lift:(1.06) lev:(0.04) [30] conv:(1.67)

48. x5=0 x9=0 529 ==> x10=0 492 <conf:(0.93)> lift:(1) lev:(-0) [-1] conv:(0.93)

49. x4=0 x6=0 514 ==> x2=0 478 <conf:(0.93)> lift:(1.06) lev:(0.03) [25] conv:(1.66)

50. x1=0 471 ==> x10=0 438 <conf:(0.93)> lift:(1) lev:(-0) [-1] conv:(0.93)

51. x4=0 x9=0 635 ==> x2=0 590 <conf:(0.93)> lift:(1.06) lev:(0.04) [30] conv:(1.65)

52. x3=0 x4=0 x10=0 564 ==> x2=0 524 <conf:(0.93)> lift:(1.06) lev:(0.03) [27] conv:(1.64)

53. x4=0 x6=0 x10=0 465 ==> x2=0 432 <conf:(0.93)> lift:(1.06) lev:(0.03) [22] conv:(1.63)

54. x2=0 x3=0 x8=0 492 ==> x10=0 457 <conf:(0.93)> lift:(1) lev:(-0) [-1] conv:(0.92)

55. x4=0 x9=0 x10=0 587 ==> x2=0 545 <conf:(0.93)> lift:(1.05) lev:(0.03) [28] conv:(1.63)

56. x2=0 708 ==> x10=0 657 <conf:(0.93)> lift:(0.99) lev:(-0) [-3] conv:(0.91)

57. x2=0 x9=0 613 ==> x10=0 568 <conf:(0.93)> lift:(0.99) lev:(-0) [-3] conv:(0.9)

58. x2=0 x5=0 517 ==> x10=0 479 <conf:(0.93)> lift:(0.99) lev:(-0) [-3] conv:(0.89)

59. x4=0 732 ==> x10=0 678 <conf:(0.93)> lift:(0.99) lev:(-0.01) [-4] conv:(0.89)

60. x3=0 x4=0 x8=0 511 ==> x10=0 473 <conf:(0.93)> lift:(0.99) lev:(-0) [-3] conv:(0.88)

61. x2=0 x4=0 684 ==> x10=0 633 <conf:(0.93)> lift:(0.99) lev:(-0.01) [-5] conv:(0.88)

62. x4=0 x5=0 536 ==> x2=0 496 <conf:(0.93)> lift:(1.05) lev:(0.03) [23] conv:(1.56)

63. x2=0 x3=0 x4=0 x8=0 469 ==> x10=0 434 <conf:(0.93)> lift:(0.99) lev:(-0) [-3] conv:(0.88)

64. x4=0 x5=0 x10=0 495 ==> x2=0 458 <conf:(0.93)> lift:(1.05) lev:(0.03) [22] conv:(1.56)

65. x4=0 x8=0 625 ==> x2=0 578 <conf:(0.92)> lift:(1.05) lev:(0.03) [27] conv:(1.55)

66. x4=0 x8=0 x10=0 585 ==> x2=0 541 <conf:(0.92)> lift:(1.05) lev:(0.03) [25] conv:(1.55)

67. x6=0 x8=0 477 ==> x10=0 441 <conf:(0.92)> lift:(0.99) lev:(-0) [-3] conv:(0.87)

68. x4=0 x9=0 635 ==> x10=0 587 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.87)

69. x2=0 x5=0 x9=0 448 ==> x10=0 414 <conf:(0.92)> lift:(0.99) lev:(-0) [-3] conv:(0.86)

70. x3=0 685 ==> x10=0 633 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.87)

71. x2=0 x4=0 x9=0 590 ==> x10=0 545 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.86)

72. x4=0 x5=0 536 ==> x10=0 495 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.86)

73. x2=0 x4=0 x5=0 496 ==> x10=0 458 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-4] conv:(0.85)

74. x3=0 x4=0 x5=0 452 ==> x2=0 417 <conf:(0.92)> lift:(1.05) lev:(0.02) [18] conv:(1.5)

75. x3=0 x5=0 516 ==> x10=0 476 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.85)

76. x3=0 x9=0 593 ==> x10=0 547 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-6] conv:(0.85)

77. x3=0 x4=0 x9=0 527 ==> x2=0 486 <conf:(0.92)> lift:(1.05) lev:(0.03) [21] conv:(1.5)

78. x4=0 x6=0 x9=0 444 ==> x2=0 409 <conf:(0.92)> lift:(1.05) lev:(0.02) [18] conv:(1.47)

79. x3=0 x4=0 x9=0 x10=0 481 ==> x2=0 443 <conf:(0.92)> lift:(1.05) lev:(0.02) [19] conv:(1.47)

80. x4=0 x5=0 x9=0 465 ==> x2=0 428 <conf:(0.92)> lift:(1.05) lev:(0.02) [18] conv:(1.46)

81. x4=0 x5=0 x9=0 465 ==> x10=0 428 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.82)

82. x3=0 x5=0 x9=0 448 ==> x10=0 412 <conf:(0.92)> lift:(0.99) lev:(-0.01) [-5] conv:(0.81)

83. x4=0 x8=0 x9=0 x10=0 503 ==> x2=0 462 <conf:(0.92)> lift:(1.04) lev:(0.02) [19] conv:(1.43)

84. x4=0 x8=0 x9=0 537 ==> x2=0 493 <conf:(0.92)> lift:(1.04) lev:(0.03) [20] conv:(1.42)

85. x3=0 x4=0 x8=0 511 ==> x2=0 469 <conf:(0.92)> lift:(1.04) lev:(0.02) [19] conv:(1.42)

86. x2=0 x3=0 596 ==> x10=0 547 <conf:(0.92)> lift:(0.98) lev:(-0.01) [-8] conv:(0.8)

87. x3=0 x4=0 x8=0 x10=0 473 ==> x2=0 434 <conf:(0.92)> lift:(1.04) lev:(0.02) [17] conv:(1.41)

88. x7=0 512 ==> x10=0 469 <conf:(0.92)> lift:(0.98) lev:(-0.01) [-8] conv:(0.78)

89. x3=0 x4=0 616 ==> x10=0 564 <conf:(0.92)> lift:(0.98) lev:(-0.01) [-10] conv:(0.78)

90. x2=0 x3=0 x9=0 508 ==> x10=0 465 <conf:(0.92)> lift:(0.98) lev:(-0.01) [-8] conv:(0.78)

91. x6=0 578 ==> x10=0 529 <conf:(0.92)> lift:(0.98) lev:(-0.01) [-10] conv:(0.78)

92. x6=0 x9=0 506 ==> x10=0 463 <conf:(0.92)> lift:(0.98) lev:(-0.01) [-9] conv:(0.77)

93. x4=0 x5=0 x8=0 458 ==> x2=0 419 <conf:(0.91)> lift:(1.04) lev:(0.02) [15] conv:(1.37)

94. x2=0 x3=0 x4=0 573 ==> x10=0 524 <conf:(0.91)> lift:(0.98) lev:(-0.01) [-10] conv:(0.77)

95. x3=0 x4=0 x9=0 527 ==> x10=0 481 <conf:(0.91)> lift:(0.98) lev:(-0.01) [-10] conv:(0.75)

96. x2=0 x3=0 x4=0 x9=0 486 ==> x10=0 443 <conf:(0.91)> lift:(0.98) lev:(-0.01) [-10] conv:(0.74)

97. x3=0 x4=0 x5=0 452 ==> x10=0 412 <conf:(0.91)> lift:(0.98) lev:(-0.01) [-9] conv:(0.74)

98. x4=0 x7=0 456 ==> x2=0 415 <conf:(0.91)> lift:(1.03) lev:(0.02) [13] conv:(1.3)

99. x3=0 x6=0 520 ==> x10=0 473 <conf:(0.91)> lift:(0.98) lev:(-0.02) [-12] conv:(0.73)

100. x3=0 x6=0 x9=0 453 ==> x10=0 412 <conf:(0.91)> lift:(0.97) lev:(-0.01) [-10] conv:(0.72)

Στο παράδειμα που παραθέτω παρακάτω:

1. x2=0 708 ==> x4=0 684 <conf:(0.97)>

Αυτοι που αγόρασαν απορρυπαντικά στην περιπτωση που εξετάζουμε υπάρχει πιθανότητα 97% να αγοράσουν και μαλακτικά

2. x2=0 x6=0 496 ==> x4=0 478 <conf:(0.96)>

Αυτοι που αγόρασαν απορρυπαντικά και προϊοντα περιποίησης στην περιπτωση που εξετάζουμε υπάρχει πιθανότητα 96% να αγοράσουν και μαλακτικά

4. x2=0 x7=0 431 ==> x4=0 415 <conf:(0.96)>

Αυτοι που αγόρασαν απορρυπαντικά και προϊοντα καθαρισμού στην περιπτωση που εξετάζουμε υπάρχει πιθανότητα 96% να αγοράσουν και μαλακτικά

4.3.2 Μέθοδος Lift

Ο κανόνας του Lift ορίζεται ως:

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(Y) * supp(X)}$$

Κατα την εκτέλεση του αλγοριθμου Lift εμφανίζεται το αποτέλεσμα για δείγμα 100 πελατών.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 1 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     semptembris-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    804
Attributes:   13
              age
              sex
              time
              x1
              x2
              x3
              x4
              x5
              x6
              x7
              x8
```

x9

x10

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.3 (241 instances)

Minimum metric <lift>: 1.1

Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 92

Size of set of large itemsets L(3): 203

Size of set of large itemsets L(4): 210

Size of set of large itemsets L(5): 98

Size of set of large itemsets L(6): 22

Size of set of large itemsets L(7): 2

Best rules found:

1. age =o 296 ==> x3=0 x6=0 252 conf:(0.85) < lift:(1.32)> lev:(0.08) [60] conv:(2.32)

2. x3=0 x6=0 520 ==> age =o 252 conf:(0.48) < lift:(1.32)> lev:(0.08) [60] conv:(1.22)

3. age =o x3=0 277 ==> x6=0 252 conf:(0.91) < lift:(1.27)> lev:(0.07) [52] conv:(2.99)

4. x6=0 578 ==> age =o x3=0 252 conf:(0.44) < lift:(1.27)> lev:(0.07) [52] conv:(1.16)

5. age =o 296 ==> x6=0 x10=0 244 conf:(0.82) < lift:(1.25)> lev:(0.06) [49] conv:(1.91)

6. x6=0 x10=0 529 ==> age =o 244 conf:(0.46) < lift:(1.25)> lev:(0.06) [49] conv:(1.17)

7. age =o 296 ==> x6=0 266 conf:(0.9) < lift:(1.25)> lev:(0.07) [53] conv:(2.68)

8. x6=0 578 ==> age =o 266 conf:(0.46) < lift:(1.25)> lev:(0.07) [53] conv:(1.17)

9. x6=0 578 ==> age =o x10=0 244 conf:(0.42) < lift:(1.24)> lev:(0.06) [47] conv:(1.14)

10. age =o x10=0 273 ==> x6=0 244 conf:(0.89) < lift:(1.24)> lev:(0.06) [47] conv:(2.56)

11. sex=m x4=0 349 ==> x2=0 x3=0 x9=0 263 conf:(0.75) < lift:(1.19)> lev:(0.05) [42] conv:(1.48)

12. x2=0 x3=0 x9=0 508 ==> sex=m x4=0 263 conf:(0.52) < lift:(1.19)> lev:(0.05) [42] conv:(1.17)

13. sex=m x2=0 338 ==> x3=0 x4=0 x9=0 263 conf:(0.78) < lift:(1.19)> lev:(0.05) [41] conv:(1.53)

14. x3=0 x4=0 x9=0 527 ==> sex=m x2=0 263 conf:(0.5) < lift:(1.19)> lev:(0.05) [41] conv:(1.15)

15. sex=m 369 ==> x2=0 x3=0 x4=0 x9=0 263 conf:(0.71) < lift:(1.18)> lev:(0.05) [39] conv:(1.36)

16. x2=0 x3=0 x4=0 x9=0 486 ==> sex=m 263 conf:(0.54) < lift:(1.18)> lev:(0.05) [39] conv:(1.17)

17. sex=m 369 ==> x3=0 x4=0 x9=0 282 conf:(0.76) < lift:(1.17)> lev:(0.05) [40] conv:(1.44)

18. x3=0 x4=0 x9=0 527 ==> sex=m 282 conf:(0.54) < lift:(1.17)> lev:(0.05) [40] conv:(1.16)

19. sex=m 369 ==> x3=0 x4=0 x9=0 x10=0 257 conf:(0.7) < lift:(1.16)> lev:(0.05) [36] conv:(1.31)

20. x3=0 x4=0 x9=0 x10=0 481 ==> sex=m 257 conf:(0.53) < lift:(1.16)> lev:(0.05) [36] conv:(1.16)

21. sex=m 369 ==> x2=0 x3=0 x9=0 271 conf:(0.73) < lift:(1.16)> lev:(0.05) [37] conv:(1.37)

22. x2=0 x3=0 x9=0 508 ==> sex=m 271 conf:(0.53) < lift:(1.16)> lev:(0.05) [37] conv:(1.15)

23. sex=m 369 ==> x2=0 x3=0 x9=0 x10=0 247 conf:(0.67) < lift:(1.16)> lev:(0.04) [33] conv:(1.26)

24. x2=0 x3=0 x9=0 x10=0 465 ==> sex=m 247 conf:(0.53) < lift:(1.16)> lev:(0.04) [33] conv:(1.15)

25. x3=0 x4=0 616 ==> x2=0 x5=0 x6=0 306 conf:(0.5) < lift:(1.14)> lev:(0.05) [38] conv:(1.12)

26. x2=0 x5=0 x6=0 349 ==> x3=0 x4=0 306 conf:(0.88) < lift:(1.14)> lev:(0.05) [38] conv:(1.85)

27. x3=0 x4=0 x9=0 527 ==> sex=m x10=0 257 conf:(0.49) < lift:(1.14)> lev:(0.04) [32] conv:(1.12)

28. sex=m x10=0 343 ==> x3=0 x4=0 x9=0 257 conf:(0.75) < lift:(1.14)> lev:(0.04) [32] conv:(1.36)

29. x2=0 x3=0 x9=0 508 ==> sex=m x10=0 247 conf:(0.49) < lift:(1.14)> lev:(0.04) [30] conv:(1.11)

30. sex=m x10=0 343 ==> x2=0 x3=0 x9=0 247 conf:(0.72) < lift:(1.14)> lev:(0.04) [30] conv:(1.3)

31. x2=0 x5=0 x6=0 349 ==> x3=0 x4=0 x9=0 260 conf:(0.74) < lift:(1.14)> lev:(0.04) [31] conv:(1.34)

32. x3=0 x4=0 x9=0 527 ==> x2=0 x5=0 x6=0 260 conf:(0.49) < lift:(1.14)> lev:(0.04) [31] conv:(1.11)

33. x2=0 x3=0 596 ==> sex=m x4=0 293 conf:(0.49) < lift:(1.13)> lev:(0.04) [34] conv:(1.11)

34. sex=m x4=0 349 ==> x2=0 x3=0 293 conf:(0.84) < lift:(1.13)> lev:(0.04) [34] conv:(1.58)

35. x3=0 x4=0 616 ==> x2=0 x5=0 x6=0 x10=0 274 conf:(0.44) < lift:(1.13)> lev:(0.04) [31] conv:(1.09)

36. x2=0 x5=0 x6=0 x10=0 316 ==> x3=0 x4=0 274 conf:(0.87) < lift:(1.13)> lev:(0.04) [31] conv:(1.72)

37. sex=m x2=0 338 ==> x3=0 x4=0 293 conf:(0.87) < lift:(1.13)> lev:(0.04) [34] conv:(1.72)

38. x3=0 x4=0 616 ==> sex=m x2=0 293 conf:(0.48) < lift:(1.13)> lev:(0.04) [34] conv:(1.1)

39. x3=0 x4=0 616 ==> x2=0 x5=0 x6=0 x9=0 260 conf:(0.42) < lift:(1.13)> lev:(0.04) [30] conv:(1.08)

40. x2=0 x5=0 x6=0 x9=0 300 ==> x3=0 x4=0 260 conf:(0.87) < lift:(1.13)> lev:(0.04) [30] conv:(1.71)

41. x2=0 x3=0 596 ==> x4=0 x5=0 x6=0 306 conf:(0.51) < lift:(1.13)> lev:(0.04) [35] conv:(1.12)

42. x4=0 x5=0 x6=0 365 ==> x2=0 x3=0 306 conf:(0.84) < lift:(1.13)> lev:(0.04) [35] conv:(1.57)

43. sex=m x2=0 338 ==> x3=0 x4=0 x10=0 268 conf:(0.79) < lift:(1.13)> lev:(0.04) [30] conv:(1.42)

44. x3=0 x4=0 x10=0 564 ==> sex=m x2=0 268 conf:(0.48) < lift:(1.13)> lev:(0.04) [30] conv:(1.1)

45. sex=m x4=0 349 ==> x2=0 x3=0 x10=0 268 conf:(0.77) < lift:(1.13)> lev:(0.04) [30] conv:(1.36)

46. x2=0 x3=0 x10=0 547 ==> sex=m x4=0 268 conf:(0.49) < lift:(1.13)> lev:(0.04) [30] conv:(1.11)

47. x4=0 x5=0 x6=0 365 ==> x2=0 x3=0 x9=0 260 conf:(0.71) < lift:(1.13)> lev:(0.04) [29] conv:(1.27)

48. x2=0 x3=0 x9=0 508 ==> x4=0 x5=0 x6=0 260 conf:(0.51) < lift:(1.13)> lev:(0.04) [29] conv:(1.11)

49. x3=0 x4=0 616 ==> x2=0 x6=0 428 conf:(0.69) < lift:(1.13)> lev:(0.06) [47] conv:(1.25)

50. x2=0 x6=0 496 ==> x3=0 x4=0 428 conf:(0.86) < lift:(1.13)> lev:(0.06) [47] conv:(1.68)

51. x2=0 x3=0 596 ==> x4=0 x5=0 x6=0 x10=0 274 conf:(0.46) < lift:(1.12)> lev:(0.04) [30] conv:(1.09)

52. x4=0 x5=0 x6=0 x10=0 329 ==> x2=0 x3=0 274 conf:(0.83) < lift:(1.12)> lev:(0.04) [30] conv:(1.52)

53. x2=0 x3=0 596 ==> x4=0 x6=0 428 conf:(0.72) < lift:(1.12)> lev:(0.06) [46] conv:(1.27)

54. x4=0 x6=0 514 ==> x2=0 x3=0 428 conf:(0.83) < lift:(1.12)> lev:(0.06) [46] conv:(1.53)

55. x3=0 x4=0 616 ==> sex=m x2=0 x9=0 263 conf:(0.43) < lift:(1.12)> lev:(0.04) [28] conv:(1.08)

56. sex=m x2=0 x9=0 306 ==> x3=0 x4=0 263 conf:(0.86) < lift:(1.12)> lev:(0.04) [28] conv:(1.63)

57. sex=m x4=0 349 ==> x2=0 x9=0 298 conf:(0.85) < lift:(1.12)> lev:(0.04) [31] conv:(1.59)

58. x2=0 x9=0 613 ==> sex=m x4=0 298 conf:(0.49) < lift:(1.12)> lev:(0.04) [31] conv:(1.1)

59. x2=0 x3=0 596 ==> sex=m x4=0 x10=0 268 conf:(0.45) < lift:(1.12)> lev:(0.04) [28] conv:(1.08)

60. sex=m x4=0 x10=0 323 ==> x2=0 x3=0 268 conf:(0.83) < lift:(1.12)> lev:(0.04) [28] conv:(1.49)

61. x2=0 x3=0 596 ==> sex=m x4=0 x9=0 263 conf:(0.44) < lift:(1.12)> lev:(0.03) [28] conv:(1.08)

62. sex=m x4=0 x9=0 317 ==> x2=0 x3=0 263 conf:(0.83) < lift:(1.12)> lev:(0.03) [28] conv:(1.49)

63. x2=0 x5=0 x6=0 349 ==> x3=0 x4=0 x10=0 274 conf:(0.79) < lift:(1.12)> lev:(0.04) [29] conv:(1.37)

64. x3=0 x4=0 x10=0 564 ==> x2=0 x5=0 x6=0 274 conf:(0.49) < lift:(1.12)> lev:(0.04) [29] conv:(1.1)

65. x3=0 685 ==> sex=m x6=0 246 conf:(0.36) < lift:(1.12)> lev:(0.03) [26] conv:(1.06)

66. sex=m x6=0 258 ==> x3=0 246 conf:(0.95) < lift:(1.12)> lev:(0.03) [26] conv:(2.94)

67. x4=0 x6=0 514 ==> x2=0 x3=0 x9=0 363 conf:(0.71) < lift:(1.12)> lev:(0.05) [38] conv:(1.24)

68. x2=0 x3=0 x9=0 508 ==> x4=0 x6=0 363 conf:(0.71) < lift:(1.12)> lev:(0.05) [38] conv:(1.26)

69. x3=0 x4=0 616 ==> sex=m x2=0 x10=0 268 conf:(0.44) < lift:(1.12)> lev:(0.04) [28] conv:(1.08)

70. sex=m x2=0 x10=0 313 ==> x3=0 x4=0 268 conf:(0.86) < lift:(1.12)> lev:(0.04) [28] conv:(1.59)

71. x2=0 x6=0 496 ==> x3=0 x4=0 x9=0 363 conf:(0.73) < lift:(1.12)> lev:(0.05) [37] conv:(1.28)

72. x3=0 x4=0 x9=0 527 ==> x2=0 x6=0 363 conf:(0.69) < lift:(1.12)> lev:(0.05) [37] conv:(1.22)

73. sex=m x2=0 338 ==> x4=0 x9=0 298 conf:(0.88) < lift:(1.12)> lev:(0.04) [31] conv:(1.73)

74. x4=0 x9=0 635 ==> sex=m x2=0 298 conf:(0.47) < lift:(1.12)> lev:(0.04) [31] conv:(1.09)

75. sex=m x4=0 349 ==> x2=0 x8=0 x9=0 250 conf:(0.72) < lift:(1.12)> lev:(0.03) [26] conv:(1.25)

76. x2=0 x8=0 x9=0 516 ==> sex=m x4=0 250 conf:(0.48) < lift:(1.12)> lev:(0.03) [26] conv:(1.09)

77. sex=m x4=0 x8=0 294 ==> x2=0 x9=0 250 conf:(0.85) < lift:(1.12)> lev:(0.03) [25] conv:(1.55)

78. x2=0 x9=0 613 ==> sex=m x4=0 x8=0 250 conf:(0.41) < lift:(1.12)> lev:(0.03) [25] conv:(1.07)

79. sex=m x2=0 x8=0 284 ==> x4=0 x9=0 250 conf:(0.88) < lift:(1.11)> lev:(0.03) [25] conv:(1.71)

80. x4=0 x9=0 635 ==> sex=m x2=0 x8=0 250 conf:(0.39) < lift:(1.11)> lev:(0.03) [25] conv:(1.06)

81. sex=m 369 ==> x2=0 x3=0 x4=0 x10=0 268 conf:(0.73) < lift:(1.11)> lev:(0.03) [27] conv:(1.26)

82. x2=0 x3=0 x4=0 x10=0 524 ==> sex=m 268 conf:(0.51) < lift:(1.11)> lev:(0.03) [27] conv:(1.1)

83. sex=m 369 ==> x2=0 x3=0 x4=0 293 conf:(0.79) < lift:(1.11)> lev:(0.04) [30] conv:(1.38)

84. x2=0 x3=0 x4=0 573 ==> sex=m 293 conf:(0.51) < lift:(1.11)> lev:(0.04) [30] conv:(1.1)

85. x2=0 x3=0 596 ==> x4=0 x6=0 x10=0 384 conf:(0.64) < lift:(1.11)> lev:(0.05) [39] conv:(1.18)

86. x4=0 x6=0 x10=0 465 ==> x2=0 x3=0 384 conf:(0.83) < lift:(1.11)> lev:(0.05) [39] conv:(1.47)

87. x3=0 x4=0 616 ==> x2=0 x6=0 x10=0 384 conf:(0.62) < lift:(1.11)> lev:(0.05) [39] conv:(1.16)

88. x2=0 x6=0 x10=0 450 ==> x3=0 x4=0 384 conf:(0.85) < lift:(1.11)> lev:(0.05) [39] conv:(1.57)

89. x2=0 x3=0 596 ==> x4=0 x5=0 x6=0 x9=0 260 conf:(0.44) < lift:(1.11)> lev:(0.03) [26] conv:(1.08)

90. x4=0 x5=0 x6=0 x9=0 315 ==> x2=0 x3=0 260 conf:(0.83) < lift:(1.11)> lev:(0.03) [26] conv:(1.46)

91. x2=0 x9=0 613 ==> sex=m x4=0 x10=0 274 conf:(0.45) < lift:(1.11)> lev:(0.03) [27] conv:(1.08)

92. sex=m x4=0 x10=0 323 ==> x2=0 x9=0 274 conf:(0.85) < lift:(1.11)> lev:(0.03) [27] conv:(1.53)

93. x3=0 685 ==> age =o x6=0 252 conf:(0.37) < lift:(1.11)> lev:(0.03) [25] conv:(1.06)

94. age =o x6=0 266 ==> x3=0 252 conf:(0.95) < lift:(1.11)> lev:(0.03) [25] conv:(2.62)

95. sex=m x4=0 349 ==> x2=0 x9=0 x10=0 274 conf:(0.79) < lift:(1.11)> lev:(0.03) [27] conv:(1.35)

96. x2=0 x9=0 x10=0 568 ==> sex=m x4=0 274 conf:(0.48) < lift:(1.11)> lev:(0.03) [27] conv:(1.09)

97. sex=m x2=0 338 ==> x4=0 x9=0 x10=0 274 conf:(0.81) < lift:(1.11)> lev:(0.03) [27] conv:(1.4)

98. x4=0 x9=0 x10=0 587 ==> sex=m x2=0 274 conf:(0.47) < lift:(1.11)> lev:(0.03) [27] conv:(1.08)

99. x3=0 x4=0 616 ==> x2=0 x6=0 x9=0 363 conf:(0.59) < lift:(1.11)> lev:(0.04) [35] conv:(1.14)

100. x2=0 x6=0 x9=0 427 ==> x3=0 x4=0 363 conf:(0.85) < lift:(1.11)> lev:(0.04) [35] conv:(1.54)

Στο παράδειμα που παραθέτω παρακάτω:

1. age =o 296 ==> x3=0 x6=0 252 < lift:(1.32)>

Αν είναι στην ηλικιακή ομάδα των μεγάλων τοτε με πιθανότητα 1.32 θα αγοράσει καλυντικά μαζί με προϊόντα περιποίησης .

33.sex=m x10=0 343 ==> x2=0 x3=0 x9=0 247 < lift:(1.14)>

Αν είναι άντρας και αγοράσει προϊόντα διακόσμησης τοτε υπαρχει πιθανότητα να αγοράσει απορρυπαντικά, καλυντικά και βρεφικά

3. age =o x3=0 277 ==> x6=0 252 < lift:(1.27)>

Αν ειναι μεγάλος και αγοράσει καλυντικά υπάρχει πιθανότητα να αγοράσει και προϊοντά περιποίησης

4.3.3 Μέθοδος Conviction

Ο κανόνας του Conviction ορίζεται ως:

$$conv(X \rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \rightarrow Y)}$$

Κατα την εκτέλεση του αλγοριθμου Conviction εμφανίζεται το αποτέλεσμα για δείγμα 100 πελατών.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 3 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     semptembris-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    804
Attributes:   13
```

```
        age

        sex

        time

        x1

        x2

        x3

        x4

        x5

        x6

        x7

        x8

        x9

        x10
=== Associator model (full training set) ===



Apriori

=======


Minimum support: 0.5 (402 instances)

Minimum metric <conviction>: 1.1

Number of cycles performed: 10


Generated sets of large itemsets:
```

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 39

Size of set of large itemsets L(3): 52

Size of set of large itemsets L(4): 32

Size of set of large itemsets L(5): 3

Best rules found:

1. x2=0 708 ==> x4=0 684 conf:(0.97) lift:(1.06) lev:(0.05) [39] < conv:(2.54)>
2. x2=0 x10=0 657 ==> x4=0 633 conf:(0.96) lift:(1.06) lev:(0.04) [34] < conv:(2.35)>
3. x2=0 x6=0 496 ==> x4=0 478 conf:(0.96) lift:(1.06) lev:(0.03) [26] < conv:(2.34)>
4. x2=0 x9=0 613 ==> x4=0 590 conf:(0.96) lift:(1.06) lev:(0.04) [31] < conv:(2.29)>
5. x2=0 x7=0 431 ==> x4=0 415 conf:(0.96) lift:(1.06) lev:(0.03) [22] < conv:(2.27)>
6. x2=0 x3=0 596 ==> x4=0 573 conf:(0.96) lift:(1.06) lev:(0.04) [30] < conv:(2.22)>
7. x2=0 x3=0 x6=0 445 ==> x4=0 428 conf:(0.96) lift:(1.06) lev:(0.03) [22] < conv:(2.21)>
8. x2=0 x8=0 602 ==> x4=0 578 conf:(0.96) lift:(1.05) lev:(0.04) [29] < conv:(2.16)>
9. x2=0 x6=0 x10=0 450 ==> x4=0 432 conf:(0.96) lift:(1.05) lev:(0.03) [22] < conv:(2.12)>
10. x2=0 x9=0 x10=0 568 ==> x4=0 545 conf:(0.96) lift:(1.05) lev:(0.03) [27] < conv:(2.12)>
11. x2=0 x5=0 517 ==> x4=0 496 conf:(0.96) lift:(1.05) lev:(0.03) [25] < conv:(2.1)>

12. x2=0 x3=0 x10=0 547 ==> x4=0 524 conf:(0.96) lift:(1.05) lev:(0.03) [25] < conv:(2.04)>

13. x2=0 x8=0 x10=0 565 ==> x4=0 541 conf:(0.96) lift:(1.05) lev:(0.03) [26] < conv:(2.02)>

14. x2=0 x6=0 x9=0 427 ==> x4=0 409 conf:(0.96) lift:(1.05) lev:(0.03) [20] < conv:(2.01)>

15. x2=0 x3=0 x9=0 508 ==> x4=0 486 conf:(0.96) lift:(1.05) lev:(0.03) [23] < conv:(1.98)>

16. x2=0 x5=0 x10=0 479 ==> x4=0 458 conf:(0.96) lift:(1.05) lev:(0.03) [21] < conv:(1.95)>

17. x2=0 x8=0 x9=0 516 ==> x4=0 493 conf:(0.96) lift:(1.05) lev:(0.03) [23] < conv:(1.93)>

18. x2=0 x5=0 x9=0 448 ==> x4=0 428 conf:(0.96) lift:(1.05) lev:(0.03) [20] < conv:(1.91)>

19. x2=0 x3=0 x5=0 437 ==> x4=0 417 conf:(0.95) lift:(1.05) lev:(0.02) [19] < conv:(1.86)>

20. x2=0 x3=0 x8=0 492 ==> x4=0 469 conf:(0.95) lift:(1.05) lev:(0.03) [21] < conv:(1.84)>

21. x2=0 x3=0 x9=0 x10=0 465 ==> x4=0 443 conf:(0.95) lift:(1.05) lev:(0.02) [19] < conv:(1.81)>

22. x2=0 x8=0 x9=0 x10=0 485 ==> x4=0 462 conf:(0.95) lift:(1.05) lev:(0.03) [20] < conv:(1.81)>

23. x2=0 x5=0 x8=0 440 ==> x4=0 419 conf:(0.95) lift:(1.05) lev:(0.02) [18] < conv:(1.79)>

24. x4=0 732 ==> x2=0 684 conf:(0.93) lift:(1.06) lev:(0.05) [39] < conv:(1.78)>

25. x4=0 x10=0 678 ==> x2=0 633 conf:(0.93) lift:(1.06) lev:(0.04) [35] < conv:(1.76)>

26. x3=0 x4=0 x6=0 459 ==> x2=0 428 conf:(0.93) lift:(1.06) lev:(0.03) [23] < conv:(1.71)>

27. x2=0 x3=0 x8=0 x10=0 457 ==> x4=0 434 conf:(0.95) lift:(1.04) lev:(0.02) [17] < conv:(1.71)>

28. x2=0 x6=0 496 ==> x3=0 x4=0 428 conf:(0.86) lift:(1.13) lev:(0.06) [47] < conv:(1.68)>

29. x3=0 x4=0 616 ==> x2=0 573 conf:(0.93) lift:(1.06) lev:(0.04) [30] < conv:(1.67)>

30. x4=0 x6=0 514 ==> x2=0 478 conf:(0.93) lift:(1.06) lev:(0.03) [25] < conv:(1.66)>

31. x4=0 x9=0 635 ==> x2=0 590 conf:(0.93) lift:(1.06) lev:(0.04) [30] < conv:(1.65)>

32. x3=0 x4=0 x10=0 564 ==> x2=0 524 conf:(0.93) lift:(1.06) lev:(0.03) [27] < conv:(1.64)>

33. x4=0 x6=0 x10=0 465 ==> x2=0 432 conf:(0.93) lift:(1.06) lev:(0.03) [22] < conv:(1.63)>

34. x4=0 x9=0 x10=0 587 ==> x2=0 545 conf:(0.93) lift:(1.05) lev:(0.03) [28] < conv:(1.63)>

35. x4=0 x5=0 536 ==> x2=0 496 conf:(0.93) lift:(1.05) lev:(0.03) [23] < conv:(1.56)>

36. x4=0 x5=0 x10=0 495 ==> x2=0 458 conf:(0.93) lift:(1.05) lev:(0.03) [22] < conv:(1.56)>

37. x4=0 x8=0 625 ==> x2=0 578 conf:(0.92) lift:(1.05) lev:(0.03) [27] < conv:(1.55)>

38. x4=0 x8=0 x10=0 585 ==> x2=0 541 conf:(0.92) lift:(1.05) lev:(0.03) [25] < conv:(1.55)>

39. x4=0 x6=0 514 ==> x2=0 x3=0 428 conf:(0.83) lift:(1.12) lev:(0.06) [46] < conv:(1.53)>

40. x2=0 x8=0 602 ==> x4=0 x10=0 541 conf:(0.9) lift:(1.07) lev:(0.04) [33] < conv:(1.52)>

41. x3=0 x4=0 x5=0 452 ==> x2=0 417 conf:(0.92) lift:(1.05) lev:(0.02) [18] < conv:(1.5)>

42. x3=0 x4=0 x9=0 527 ==> x2=0 486 conf:(0.92) lift:(1.05) lev:(0.03) [21] < conv:(1.5)>

43. x3=0 x4=0 x9=0 x10=0 481 ==> x2=0 443 conf:(0.92) lift:(1.05) lev:(0.02) [19] < conv:(1.47)>

44. x4=0 x6=0 x9=0 444 ==> x2=0 409 conf:(0.92) lift:(1.05) lev:(0.02) [18] < conv:(1.47)>

45. x2=0 x8=0 x9=0 516 ==> x4=0 x10=0 462 conf:(0.9) lift:(1.06) lev:(0.03) [26] < conv:(1.47)>

46. x4=0 x5=0 x9=0 465 ==> x2=0 428 conf:(0.92) lift:(1.05) lev:(0.02) [18] < conv:(1.46)>

47. x2=0 708 ==> x4=0 x10=0 633 conf:(0.89) lift:(1.06) lev:(0.04) [35] < conv:(1.46)>

48. x6=0 578 ==> x3=0 520 conf:(0.9) lift:(1.06) lev:(0.03) [27] < conv:(1.45)>

49. x4=0 x8=0 x9=0 x10=0 503 ==> x2=0 462 conf:(0.92) lift:(1.04) lev:(0.02) [19] < conv:(1.43)>

50. x4=0 x8=0 x9=0 537 ==> x2=0 493 conf:(0.92) lift:(1.04) lev:(0.03) [20] < conv:(1.42)>

51. x3=0 x4=0 x8=0 511 ==> x2=0 469 conf:(0.92) lift:(1.04) lev:(0.02) [19] < conv:(1.42)>

52. x3=0 x4=0 x8=0 x10=0 473 ==> x2=0 434 conf:(0.92) lift:(1.04) lev:(0.02) [17] < conv:(1.41)>

53. x2=0 x6=0 496 ==> x3=0 445 conf:(0.9) lift:(1.05) lev:(0.03) [22] < conv:(1.41)>

54. x2=0 x9=0 613 ==> x4=0 x10=0 545 conf:(0.89) lift:(1.05) lev:(0.03) [28] < conv:(1.39)>

55. x2=0 x4=0 x6=0 478 ==> x3=0 428 conf:(0.9) lift:(1.05) lev:(0.03) [20] < conv:(1.39)>

56. x6=0 x9=0 506 ==> x3=0 453 conf:(0.9) lift:(1.05) lev:(0.03) [21] < conv:(1.39)>

57. x6=0 x10=0 529 ==> x3=0 473 conf:(0.89) lift:(1.05) lev:(0.03) [22] < conv:(1.37)>

58. x4=0 x5=0 x8=0 458 ==> x2=0 419 conf:(0.91) lift:(1.04) lev:(0.02) [15] < conv:(1.37)>

59. x4=0 x6=0 514 ==> x3=0 459 conf:(0.89) lift:(1.05) lev:(0.03) [21] < conv:(1.36)>

60. x2=0 x5=0 517 ==> x4=0 x10=0 458 conf:(0.89) lift:(1.05) lev:(0.03) [22] < conv:(1.35)>

61. x4=0 x8=0 625 ==> x2=0 x10=0 541 conf:(0.87) lift:(1.06) lev:(0.04) [30] < conv:(1.34)>

62. x4=0 732 ==> x2=0 x10=0 633 conf:(0.86) lift:(1.06) lev:(0.04) [34] < conv:(1.34)>

63. x6=0 x9=0 x10=0 463 ==> x3=0 412 conf:(0.89) lift:(1.04) lev:(0.02) [17] < conv:(1.32)>

64. x2=0 x3=0 x8=0 492 ==> x4=0 x10=0 434 conf:(0.88) lift:(1.05) lev:(0.02) [19] < conv:(1.31)>

65. x4=0 x7=0 456 ==> x2=0 415 conf:(0.91) lift:(1.03) lev:(0.02) [13] < conv:(1.3)>

66. x4=0 x8=0 x9=0 537 ==> x2=0 x10=0 462 conf:(0.86) lift:(1.05) lev:(0.03) [23] < conv:(1.29)>

67. x2=0 x3=0 596 ==> x4=0 x10=0 524 conf:(0.88) lift:(1.04) lev:(0.03) [21] < conv:(1.28)>

68. x4=0 x9=0 635 ==> x2=0 x10=0 545 conf:(0.86) lift:(1.05) lev:(0.03) [26] < conv:(1.28)>

69. x4=0 x6=0 x10=0 465 ==> x3=0 412 conf:(0.89) lift:(1.04) lev:(0.02) [15] < conv:(1.27)>

70. x2=0 x3=0 596 ==> x4=0 x6=0 428 conf:(0.72) lift:(1.12) lev:(0.06) [46] < conv:(1.27)>

71. x2=0 708 ==> x4=0 x9=0 590 conf:(0.83) lift:(1.06) lev:(0.04) [30] < conv:(1.25)>

72. x2=0 x10=0 657 ==> x4=0 x8=0 541 conf:(0.82) lift:(1.06) lev:(0.04) [30] < conv:(1.25)>

73. x3=0 x4=0 616 ==> x2=0 x6=0 428 conf:(0.69) lift:(1.13) lev:(0.06) [47] < conv:(1.25)>

74. x4=0 x5=0 536 ==> x2=0 x10=0 458 conf:(0.85) lift:(1.05) lev:(0.02) [20] < conv:(1.24)>

75. x6=0 x8=0 477 ==> x3=0 421 conf:(0.88) lift:(1.04) lev:(0.02) [14] < conv:(1.24)>

76. x4=0 x10=0 678 ==> x2=0 x8=0 541 conf:(0.8) lift:(1.07) lev:(0.04) [33] < conv:(1.23)>

77. x2=0 x10=0 657 ==> x4=0 x9=0 545 conf:(0.83) lift:(1.05) lev:(0.03) [26] < conv:(1.22)>

78. x2=0 708 ==> x3=0 x4=0 573 conf:(0.81) lift:(1.06) lev:(0.04) [30] < conv:(1.22)>

79. x4=0 732 ==> x2=0 x9=0 590 conf:(0.81) lift:(1.06) lev:(0.04) [31] < conv:(1.22)>

80. x3=0 x4=0 616 ==> x2=0 x10=0 524 conf:(0.85) lift:(1.04) lev:(0.03) [20] < conv:(1.21)>

81. x2=0 x5=0 517 ==> x4=0 x9=0 428 conf:(0.83) lift:(1.05) lev:(0.02) [19] < conv:(1.21)>

82. x2=0 x3=0 x9=0 508 ==> x4=0 x10=0 443 conf:(0.87) lift:(1.03) lev:(0.02) [14] < conv:(1.21)>

83. x6=0 578 ==> x3=0 x9=0 453 conf:(0.78) lift:(1.06) lev:(0.03) [26] < conv:(1.2)>

84. x2=0 708 ==> x4=0 x8=0 578 conf:(0.82) lift:(1.05) lev:(0.03) [27] < conv:(1.2)>

85. x4=0 x10=0 678 ==> x2=0 x9=0 545 conf:(0.8) lift:(1.05) lev:(0.03) [28] < conv:(1.2)>

86. x3=0 x4=0 x8=0 511 ==> x2=0 x10=0 434 conf:(0.85) lift:(1.04) lev:(0.02) [16] < conv:(1.2)>

87. x2=0 x5=0 517 ==> x3=0 x4=0 417 conf:(0.81) lift:(1.05) lev:(0.03) [20] < conv:(1.2)>

88. x2=0 x6=0 496 ==> x4=0 x10=0 432 conf:(0.87) lift:(1.03) lev:(0.02) [13] < conv:(1.2)>

89. x4=0 732 ==> x2=0 x8=0 578 conf:(0.79) lift:(1.05) lev:(0.04) [29] < conv:(1.19)>

90. x2=0 x6=0 496 ==> x4=0 x9=0 409 conf:(0.82) lift:(1.04) lev:(0.02) [17] < conv:(1.18)>

91. x4=0 732 ==> x2=0 x3=0 573 conf:(0.78) lift:(1.06) lev:(0.04) [30] < conv:(1.18)>

92. x3=0 x9=0 593 ==> x6=0 453 conf:(0.76) lift:(1.06) lev:(0.03) [26] < conv:(1.18)>

93. x2=0 x9=0 x10=0 568 ==> x4=0 x8=0 462 conf:(0.81) lift:(1.05) lev:(0.03) [20] < conv:(1.18)>

94. x6=0 x10=0 529 ==> x3=0 x9=0 412 conf:(0.78) lift:(1.06) lev:(0.03) [21] < conv:(1.18)>

95. x4=0 x9=0 x10=0 587 ==> x2=0 x8=0 462 conf:(0.79) lift:(1.05) lev:(0.03) [22] < conv:(1.17)>

96. x4=0 x5=0 536 ==> x2=0 x9=0 428 conf:(0.8) lift:(1.05) lev:(0.02) [19] < conv:(1.17)>

97. x2=0 708 ==> x4=0 x9=0 x10=0 545 conf:(0.77) lift:(1.05) lev:(0.03) [28] < conv:(1.17)>

98. x2=0 x5=0 517 ==> x4=0 x8=0 419 conf:(0.81) lift:(1.04) lev:(0.02) [17] < conv:(1.16)>

99. x3=0 685 ==> x6=0 520 conf:(0.76) lift:(1.06) lev:(0.03) [27] < conv:(1.16)>

100. x6=0 578 ==> x3=0 x10=0 473 conf:(0.82) lift:(1.04) lev:(0.02) [17] < conv:(1.16)>

Στο παράδειμα που παραθέτω παρακάτω:

1. x2=0 708 ==> x4=0 684 < conv:(2.54)>

Αυτός που θα αγοράσει απορρυπαντικά θα αγοράσει μαλλακτικά με πιθανότητα 2.54

2. x2=0 x10=0 657 ==> x4=0 633 < conv:(2.35)>

Αυτός που θα αγοράσει απορρυπαντικά και προϊόντα διακόσμησης θα αγοράσει και μαλλακτικά με πιθανότητα 2.35

3. x2=0 x6=0 496 ==> x4=0 478 < conv:(2.34)>

Αυτος που θα αγοράσει απορρυπαντικά και προϊόντα περιποίησης θα αγοράσει και μαλλακτικά με πιθανότητα 2.34

4.3.4 Μέθοδος Leverage

Ο κανόνας του Leverage ορίζεται ως:

$$PS(X \Rightarrow Y) = leverage(X \Rightarrow Y) = supp(X \Rightarrow Y) - supp(X)supp(Y) = P(X \cap Y) - P(X)P(Y)$$

Κατα την εκτέλεση του αλγοριθμου Leverage εμφανίζεται το αποτέλεσμα για δείγμα 100 πελατών.

=== Run information ===

Scheme: weka.associations.Apriori -N 100 -T 2 -C 0.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: semptembris-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 804

Attributes: 13

age

sex

time

x1

x2

x3

x4

x5

x6

x7

x8

x9

x10

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (80 instances)

Minimum metric <leverage>: 0.1

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 22

Size of set of large itemsets L(2): 154

Size of set of large itemsets L(3): 600

Size of set of large itemsets L(4): 1430

Size of set of large itemsets L(5): 1953

Size of set of large itemsets L(6): 1391

Size of set of large itemsets L(7): 485

Size of set of large itemsets L(8): 76

Size of set of large itemsets L(9): 3

Best rules found:

Βιβλιογραφία

[1]. W. Frawley and G. Piatetsky-Shapiro and C. Matheus (Fall 1992). "Knowledge Discovery in Databases: An Overview". AI Magazine: pp. 213-228. ISSN 0738-4602.

[2]. D. Hand, H. Mannila, P. Smyth (2001). Principles of Data Mining. MIT Press, Cambridge.MA

[3]. Ellen Monk, Bret Wagner (2006). Concepts in Enterprise Resource Planning, Second Edition. Thomson Course Technology, Boston, MA. ISBN 0-619-21663-8. OCLC 224465825.

[4]. Albion Research, Market Basket Analysis. http://www.albionresearch.com/data_mining/market_basket.php

[5]. Gayle S., The Marriage of Market Basket Analysis to Predictive Modeling

[6]. Agrawal R., Imielinski T., Swami A. Mining Association Rules Between Sets of Items in Large Databases

[7]. D. Holmes “Authorship attribution” Computers and the Humanities 28 (1994), 87-106.

[8]. D. Holmes “The Evolution of Stylometry in Humanities Scholarship” Literary and Linguistic Computing 13 (1998),111117.http://llc.oxfordjournals.org/cgi/reprint/13/3/111.pdf

[9]. T. McEnery & M. Oates “Authorship identification and computational stylometry” in Dale et al (eds) Handbook of Natural Language Processing, New York (2000): Dekker, chapter 23.1

[10]. F. Mosteller and D. Wallace (1964). Inference and Disputed Authorship: The Federalist. Reading, MA: Addison-Wesley.

[11]. http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf

[12]. http://www.icaen.uiowa.edu/~comp/Public/Apriori.pdf

[13]. http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput499/slides/Lect10 /sld044.htm

[14]. Gionis, H. Mannila, T. Mielikainen, and P. Tsaparas, Assessing Data Mining Results via Swap Randomization, ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 1 , Issue 3 (December 2007) Article No. 14.

[15]. Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery 8:53-87, 2004.

[16]. R. Agrawal, H. Mannila, R. Srikant, A.I. Verkamo, “Fast Discovery of Association Rules”, in 3., pp. 307-328, 1996.

[17]. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “Advances in Knowledge Discovery and Data Mining”, AAAI Press/MIT Press, 1996.

[18]. Hearst Marti A. “untangling Text Data Mining”, Proceedings of ACL’99:the 37th Annual meeting of the association for computational Linguistics, University of Maryland, June 20-26,1999.

[19]. Karanikas H., Theodoulidis B. “Knowledge Discovery in text and text mining software”. Centre for Research in information management: November 2002.

[20]. Mooney R.J and Nahm Un Yong, Text Mining with information extraction. Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium,22-23 September 2003.

[21]. Besancon R & Rajman M (1998). Text Mining knowledge Extraction from unstructured textual data. 6th conference of international Federation Of classification Societies, Rome

[22]. Sehgal, A.K Text Mining: The search for novelty in text Ph.D.Comprehensive Examination Report, Dept. of Computer Science, The university of Iowa, April 2004

[23]. Albion Research, Market Basket Analysis. http://www.albionresearch.com/data_mining/market_basket.php

[24]. Data Mining, wikipedia http://en.wikipedia.org/wiki/Data_mining

[25]. Fayyad U., Piatetsky-Shapiro G., and Smyth P., From Data Mining to Knowledge Discovery in Databases.

[26]. Gayle S., The Marriage of Market Basket Analysis to Predictive Modeling

[27]. Hamilton H, Gurak E., Findlater L., Olive W., and Ranson J. Knowledge Discovery in Databases, Department of Computer Science, University of Regina, http://www2.cs.uregina.ca/%7Ehamilton/courses/831/index.html

[28]. Han J. , Kamber M. Data Mining: Concepts and Techniques

[29]. Palace B., Data Mining: What is Data Mining? Anderson Graduate School of Management at UCLA. http://www.anderson.ucla.edu/faculty/jason.frand/ teacher/technologies/palace/datamining.htm

[30]. Thearling K. An Introduction to Data Mining http://www.thearling.com/text/ dmwhite/dmwhite.htm

[31]. Yibin S., Hamilton H, Liu M., Apriori Implementation, University of Regina and Su Yi

[32].