ΣΧΟΛΗ : ΣΔΟ
ΤΜΗΜΑ : ΛΟΓΙΣΤΙΚΗΣ

ΑΤΕΙ ΠΑΤΡΑΣ 2005

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

'ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ ΠΑΝΩ ΣΕ ΘΕΜΑΤΑ ΛΟΓΙΣΤΙΚΗΣ'

ΓΕΩΡΓΙΑΔΟΥ ΕΛΕΝΗ
ΖΟΥΠΑΝΟΣ ΜΙΧΑΗΛ
ΖΗΚΑ ΣΤΑΥΡΟΥΛΑ

ΕΙΣΗΓΗΤΗΣ : ΚΩΤΣΙΑΝΤΗΣ ΣΩΤΗΡΗΣ

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Οι κανόνες συσχέτισης (association rules) αποτελούν μία σχετικά σύγχρονη μέθοδο για την εξαγωγή γνώσης από μεγάλες βάσεις δεδομένων, καθότι πρωτοεμφανίστηκε το 1993.

Οι πληροφορίες που μπορούν να περιγράψουν και να συγκεντρώσουν οι κανόνες συσχέτισης είναι ιδιαίτερα σημαντικές και αφορούν στους διάφορους τομείς της ζωής και ενασχόλησης του ανθρώπου. Κάτι τέτοιο εξάλλου αντικατοπτρίζεται και από το γεγονός ότι έχει γίνει μία σημαντική μελέτη στο πεδίο αυτό, τα τελευταία χρόνια και έχουν αναπτυχθεί πληθώρα αλγορίθμων που παράγουν κανόνες συσχέτισης.

Οι κανόνες συσχέτισης μπορούμε να πούμε ότι εμφανίστηκαν για τις ανάγκες της ανάλυσης του καλαθιού της νοικοκυράς (market basket analysis). Ο όρος αυτός προέρχεται από τις αγορές 'super-markets' στις οποίες ο καταναλωτής τοποθετεί σε ένα καλάθι το σύνολο των προϊόντων που επιθυμεί να αγοράσει. Οι υπεραγορές αυτές συγκεντρώνουν ένα τεράστιο όγκο πληροφοριών σχετικά με τις αγορές των πελατών τους, καθώς οι συναλλαγές κάθε πελάτη μπορούν να καταχωρηθούν πλέον ηλεκτρονικά. Έτσι δημιουργήθηκε η ιδέα της αξιοποίησης αυτής της πληροφορίας. Οι κανόνες συσχέτισης απλά εκφράζουν το αποτέλεσμα της ανάλυσης των χιλιάδων καλαθιών αγοράς των πελατών.

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

Οι κανόνες συσχέτισης βρίσκουν πεδίο εφαρμογής σε διάφορες πτυχές της καθημερινότητας. Στην συγκεκριμένη πτυχιακή χρησιμοποιήθηκαν για εξαγωγή χρήσιμων συμπερασμάτων για τις ηλεκρονικές πληρωμές (μέσω e-banking).

Η δομή της πτυχιακής είναι η ακόλουθη. Στο πρώτο κεφάλαιο αναφέρουμε τις εισαγωγικές έννοιες της εξόρυξης γνώσης, επιστήμης στην οποία υπάγονται οι κανόνες συσχέτισης. Στο δεύτερο κεφάλαιο αναλύεται η κυριότερη εφαρμογή των κανόνων συσχέτισης: η ανάλυση του καλαθιού της νοικοκυράς. Στο τρίτο κεφάλαιο αναφέρονται οι πιο γνωστοί αλγόριθμοι παραγωγής κανόνων συσχέτισης. Στο τέταρτο κεφάλαιο περιγράφεται το εργαλείο «DB2 INTELLIGENT MINER FOR DATA» το οποίο υλοποιεί αλγόριθμους παραγωγής κανόνων συσχέτισης και χρησιμοποιήθηκε για τα πειράματα μας. Το πέμπτο κεφάλαιο περιγράφει τα αποτελέσματα που προέκυψαν από τη χρήση του εργαλείου σε διαθέσιμα δεδομένα από ηλεκρονικές πληρωμές (μέσω e-banking). Τέλος, στο τελευταίο κεφάλαιο αναφέρονται τα συμπεράσματα της παρούσας πτυχιακής εργασίας.

1. ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΜΕΣΑ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

1.1 Εισαγωγή

Τα τελευταία χρόνια με την εξάπλωση της χρήσης των υπολογιστών σε όλους τους τομείς της ζωής μας έχουν αυξηθεί σημαντικά οι δυνατότητές μας να παράγουμε και να συλλέγουμε πληροφορίες, γεγονός που οδήγησε στην συγκέντρωση μεγάλου όγκου πληροφορίας. Η αύξηση αυτή κάνει επιτακτική την ανάγκη εύρεσης νέων τεχνικών και εργαλείων που θα υποστηρίζουν την αυτόματη μετατροπή των υπό επεξεργασία πληροφοριών σε χρήσιμη γνώση.

Ένα νέο πεδίο έρευνας, η Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων *(Knowledge Discovery in Databases - KDD)* έχει καθιερωθεί ως το κατεξοχήν πεδίο για την ανακάλυψη κρίσιμων πληροφοριών που ενδέχεται να υπάρχουν «κρυμμένες» μέσα σε μεγάλες βάσεις δεδομένων.

Η ανακάλυψη γνώσης αποτελεί έναν ταχέως αναπτυσσόμενο τομέα, η εξέλιξη του οποίου κατευθύνεται τόσο από ερευνητικά ενδιαφέροντα όσο και από ισχυρές πρακτικές, οικονομικές και κοινωνικές ανάγκες. Στο κεφάλαιο αυτό θα παρουσιάσουμε τον ορισμό και τις βασικές έννοιες και μεθόδους που χρησιμοποιούνται στην Ανακάλυψη Γνώσης.

1.2 Ανακάλυψη Γνώσης

Θα δώσουμε ένα γενικό ορισμό της έννοιας της ανακάλυψης γνώσης μέσα από βάσεις δεδομένων.

Ορισμός

Η **Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων** *(KDD)* είναι μια μη-τετριμμένη διαδικασία για την αναγνώριση έγκυρων, νέων, χρήσιμων και εύκολα κατανοητών προτύπων από τα δεδομένα.

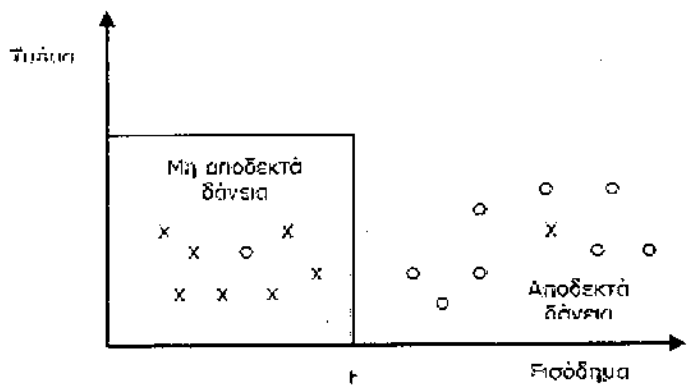
Ας δούμε όμως αναλυτικά τι σημαίνει κάθε επιμέρους όρος του ορισμού:

- Δεδομένα

Πρόκειται για ένα σύνολο παραδειγμάτων / στιγμιότυπων ενός προβλήματος που εμφανίζονται σε μια βάση δεδομένων. Για παράδειγμα, θα μπορούσε να είναι μια συλλογή εγγραφών από τη βάση δεδομένων μιας τράπεζας, όπου κάθε εγγραφή θα περιλάμβανε τρία πεδία (γνωρίσματα): το χρέος, το εισόδημα και την κατάσταση του δανείου των πελατών της τράπεζας.

- Πρότυπα *(patterns)*

Πρόκειται για εκφράσεις σε μια συγκεκριμένη γλώσσα οι οποίες περιγράφουν ένα υποσύνολο του συνόλου των παραδειγμάτων. Για παράδειγμα, η έκφραση '*Αν το εισόδημα<t, τότε ο πελάτης δεν μπορεί να εξοφλήσει το δάνειο*', θα μπορούσε να είναι ένα πρότυπο για κάποιο κατάλληλο κατώφλι *t* (Σχήμα 1).



Σχήμα 1. Ταξινόμηση των δεδομένων με χρήση κατάλληλου κατωφλίου *t*. Η σκιασμένη περιοχή του σχήματος είναι η περιοχή των μη αποδεκτών δανείων.

- KDD Διαδικασία

Πρόκειται για μια διαδικασία πολλών βημάτων που περιλαμβάνει την κατάλληλη προετοιμασία των δεδομένων, την αναζήτηση προτύπων και την αξιολόγηση της αποκτηθείσας γνώσης. Η KDD διαδικασία δεν είναι τετριμμένη καθώς εμπεριέχει κάποιο βαθμό αυτονομίας. Στο παράδειγμα του δανείου που αναφέραμε πιο πριν, ο υπολογισμός του μέσου όρου εισοδήματος του πελάτη αποτελεί πολύ χρήσιμο αποτέλεσμα, σε καμία όμως περίπτωση δεν αποτελεί ανακάλυψη γνώσης.

- Εγκυρότητα

Τα πρότυπα που προκύπτουν από τη διαδικασία ανακάλυψης γνώσης θα πρέπει να ισχύουν με κάποιο βαθμό βεβαιότητας και για νέα, άγνωστα στιγμιότυπα του

προβλήματος. Για παράδειγμα, αν στο πρότυπο που απεικονίζεται στο Σχήμα 1 το κατώφλι μετακινηθεί προς τα δεξιά τότε το μέτρο βεβαιότητας θα μειωθεί καθώς περισσότερα αποδεκτά μέχρι πρότινος δάνεια θα ανήκουν πλέον στην περιοχή των μη αποδεκτών δανείων.

- Χρησιμότητα

Τα πρότυπα θα πρέπει να είναι χρήσιμα, δηλαδή να οδηγούν σε κάποιες χρήσιμες ενέργειες. Για παράδειγμα, αν η τράπεζα εκμεταλλευτεί τους κανόνες απόφασης του Σχήματος 1, θα πρέπει να πετύχει αύξηση των κερδών της.

- Κατανοησιμότητα

Τα πρότυπα θα πρέπει να είναι κατανοητά από τον ανθρώπινο παράγοντα, καθώς οι άνθρωποι είναι αυτοί που θα κληθούν να τα αξιοποιήσουν προκειμένου να εξάγουν χρήσιμα συμπεράσματα και να αποκτήσουν μια βαθύτερη κατανόηση των δεδομένων τους. Για την κατανόηση των προτύπων δε θα πρέπει να απαιτούνται εξειδικευμένες γνώσεις, αντιθέτως τα πρότυπα θα πρέπει να είναι πλήρως κατανοητά και να βοηθούν ακόμη και μη ειδικούς στην εξαγωγή χρήσιμων συμπερασμάτων.

Η Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων (KDD) αναφέρεται στη συνολική διαδικασία ανακάλυψης χρήσιμης πληροφορίας από τα δεδομένα. Ένα βήμα σ' αυτή τη διαδικασία αποτελεί και η Εξόρυξη Γνώσης *(Data Mining)*, της οποίας ο ορισμός ακολουθεί.

Ορισμός

Η **Εξόρυξη Γνώσης** αποτελεί ένα βήμα της KDD διαδικασίας και ορίζεται ως η διαδικασία της ανακάλυψης νέων πιθανώς κρυμμένων προτύπων και μοντέλων με αυτόματο ή ημιαυτόματο τρόπο με απώτερο στόχο την περιγραφή των δεδομένων μιας βάσης δεδομένων και την πρόβλεψη και εξήγηση νέων δεδομένων.

Η Εξόρυξη Γνώσης περιλαμβάνει κυρίως τις διαδικασίες και τα μέσα εξαγωγής προτύπων από το σύνολο των δεδομένων. Μέχρι πρόσφατα αφορούσε αποκλειστικά δομημένα δεδομένα (δηλαδή δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων), τα τελευταία όμως χρόνια το ενδιαφέρον στράφηκε και σε μη δομημένα δεδομένα (π.χ. κείμενα, εικόνες, έγγραφα, web σελίδες).

1.3 Η KDD διαδικασία

Η KDD διαδικασία είναι μια αλληλεπιδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων στα οποία χρειάζεται πολλές φορές να παρέμβει και ο άνθρωπος λαμβάνοντας κρίσιμες αποφάσεις.

Τα βασικά βήματα της KDD διαδικασίας είναι τα ακόλουθα (Σχήμα 2):

- *Ανάπτυξη και κατανόηση του πεδίου της εφαρμογής* συμπεριλαμβανόμενης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα καθώς επίσης και των στόχων / προσδοκιών των τελικών χρηστών.

- *Δημιουργία του στοχευόμενου συνόλου δεδομένων (target data), το οποίο θα περιλαμβάνει* τα δεδομένα από τα οποία πρόκειται να εξαχθεί η γνώση. Το βήμα αυτό είναι εξαιρετικά κρίσιμο καθώς η ποιότητα των δεδομένων επηρεάζει την απόδοση του συστήματος ανακάλυψης γνώσης.
- *Καθαρισμός και επεξεργασία των δεδομένων (data cleaning)*. Το βήμα αυτό περιλαμβάνει βασικές λειτουργίες όπως η απομάκρυνση του θορύβου, η αντιμετώπιση του προβλήματος των δεδομένων με ελλιπείς τιμές κ.α.

- *Μείωση της ποσότητας των δεδομένων (data reduction)*. Το βήμα αυτό περιλαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων του προβλήματος ανάλογα με τους στόχους της ανακάλυψης γνώσης, τη μείωση του πλήθους αυτών των χαρακτηριστικών κ.α.

- *Επιλογή των εργασιών εξόρυξης γνώσης (data mining)* που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος, π.χ. ταξινόμηση, πρόβλεψη, ομαδοποίηση κ.α.
- *Επιλογή των αλγορίθμων εξόρυξης γνώσης (data mining)* που θα χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Το βήμα αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου, την επιλογή των κατάλληλων παραμέτρων του μοντέλου κ.α.

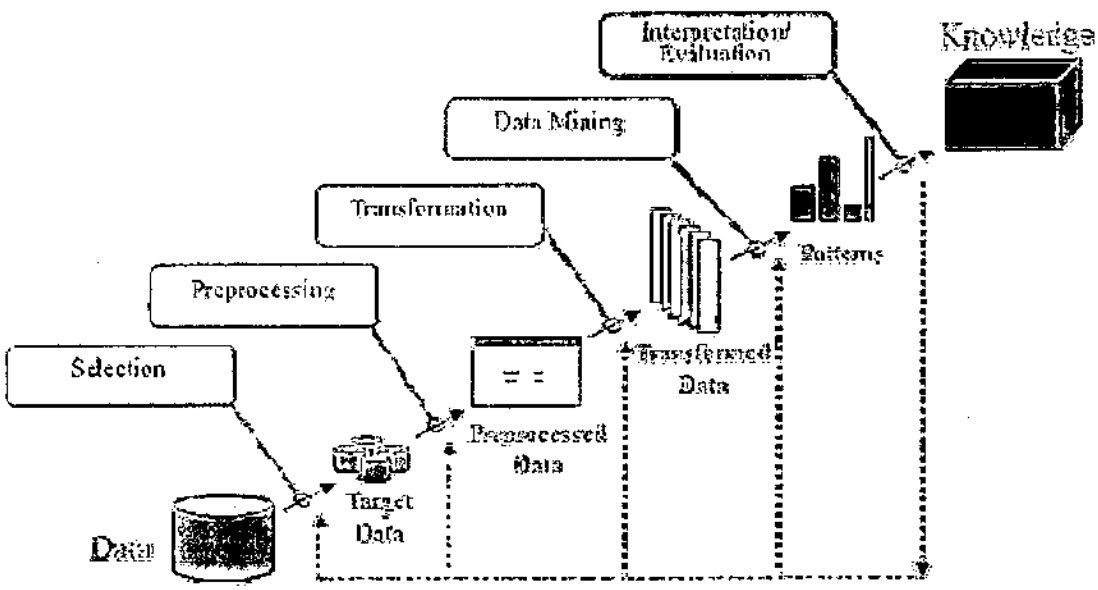
- *Data Mining*: αναζήτηση στα δεδομένα των προτύπων που μας

ενδιαφέρουν.

- *Ερμηνεία των προτύπων* που ανακαλύφθηκαν από την KDD διαδικασία - πιθανόν να χρειαστεί να επιστρέψουμε και πάλι σε κάποιο από τα παραπάνω βήματα.

- *Ενοποίηση της γνώσης που έχει εξαχθεί:* ενσωμάτωση αυτής της γνώσης στο σύστημα ή απλά κοινοποίησή της με την κατάλληλη τεκμηρίωση στα ενδιαφερόμενα μέλη. Το βήμα αυτό περιλαμβάνει και έλεγχο συγκρούσεων με την γνώση που επικρατούσε πριν.

Από τα διάφορα βήματα της KDD διαδικασίας αυτό που συγκεντρώνει το μεγαλύτερο ενδιαφέρον είναι το βήμα του data mining. Αυτό όμως, δε σημαίνει πως τα υπόλοιπα βήματα δεν είναι σημαντικά, αντιθέτως η επιτυχής τους διεκπεραίωση επηρεάζει την επιτυχία ολόκληρης της KDD διαδικασίας.



Σχήμα 2. Τα βήματα της KDD διαδικασίας.

1.4 Εξόρυξη γνώσης

Η εξόρυξη γνώσης *(data mining)* αναφέρεται στην εξαγωγή προτύπων από τα

εξεταζόμενα δεδομένα ή στην προσαρμογή ήδη υπαρχόντων μοντέλων στα δεδομένα αυτά. Τα μοντέλα παίζουν το ρόλο της γνώσης που εξάγεται από το σύνολο των δεδομένων. Η απόφαση για το αν τα μοντέλα αυτά αντανακλούν ή όχι χρήσιμη γνώση είναι μέρος της συνολικής KDD διαδικασίας και συνήθως λαμβάνεται από κάποιον ανθρώπινο παράγοντα.

Οι βασικοί στόχοι του data mining είναι η πρόβλεψη και η περιγραφή.

- Η **πρόβλεψη** *(prediction)* αναφέρεται στην πρόβλεψη της τιμής κάποιου συγκεκριμένου γνωρίσματος ενός προβλήματος και αφορά νέα στιγμιότυπα του προβλήματος.
- Η **περιγραφή** *(description)* αναφέρεται στην εύρεση εύκολα διερμηνεύσιμων προτύπων από τα δεδομένα κάποιου προβλήματος. Τα πρότυπα αυτά θα πρέπει να αποτελούν στην ουσία συμπαγείς και περιεκτικές αναπαραστάσεις των δεδομένων του προβλήματος.

Σήμερα υπάρχει πληθώρα αλγορίθμων οι οποίοι προέρχονται από διαφορετικά επιστημονικά πεδία όπως: Στατιστική, Αναγνώριση Προτύπων, Μηχανική Μάθηση, Βάσεις Δεδομένων κ.α. Παρά τη διαφορετικότητα των πεδίων, οι αλγόριθμοι αυτοί χαρακτηρίζονται από τα ακόλουθα κοινά στοιχεία:

- **το μοντέλο**

Υπάρχουν δύο παράγοντες που σχετίζονται με το μοντέλο:

- *Η λειτουργία του μοντέλου*, η οποία καθορίζει τις βασικές εργασίες που θα διεκπεραιωθούν κατά τη διάρκεια του data mining, π.χ. ταξινόμηση, ομαδοποίηση κ.α.
- *Ο τύπος αναπαράστασης του μοντέλου*, ο οποίος καθορίζει τόσο την προσαρμοστικότητα του μοντέλου στην αναπαράσταση των δεδομένων όσο και τη δυνατότητα ερμηνείας του μοντέλου με όρους κατανοητούς από τον άνθρωπο. Τυπικά, τα πιο πολύπλοκα μοντέλα προσαρμόζονται καλύτερα στα δεδομένα, αλλά ενδέχεται να είναι πιο δύσκολο να γίνουν κατανοητά και να προσαρμοστούν σε πραγματικά δεδομένα. Οι πιο γνωστές αναπαραστάσεις μοντέλων είναι τα δέντρα απόφασης, οι κανόνες, τα γραμμικά μοντέλα, τα γραφικά μοντέλα που

βασίζονται σε πιθανότητες, τα νευρωνικά δίκτυα κ.ο.κ.

- **την αξιολόγηση του μοντέλου**

Η αξιολόγηση, η οποία γίνεται βάσει κάποιων κριτηρίων αξιολόγησης (π.χ. *maximum likelihood*), καθορίζει κατά πόσο ένα συγκεκριμένο μοντέλο και οι παράμετροι του προσαρμόζονται στα κριτήρια της KDD διαδικασίας. Η αξιολόγηση ενός μοντέλου περιλαμβάνει τόσο την εκτίμηση της εγκυρότητας των προτύπων που παράγονται από αυτό όσο και την εκτίμηση της ακρίβειας, της χρησιμότητας και της ευκολίας κατανόησης του μοντέλου.

- **τον αλγόριθμο αναζήτησης**

Αναφέρεται στον καθορισμό ενός αλγορίθμου για την εύρεση συγκεκριμένων μοντέλων και παραμέτρων, με βάση ένα σύνολο δεδομένων, μια οικογένεια μοντέλων και ένα κριτήριο αξιολόγησης. Οι αλγόριθμοι αναζήτησης χωρίζονται σε δύο τύπους:

- *Αλγόριθμοι αναζήτησης παραμέτρων*, οι οποίοι αναζητούν τις παραμέτρους εκείνες που θα βελτιστοποιήσουν το μοντέλο ως προς το κριτήριο αξιολόγησης. Εκτελούν την αναζήτηση λαμβάνοντας ως είσοδο το σύνολο των δεδομένων και την αναπαράσταση του μοντέλου.
- *Αλγόριθμοι αναζήτησης μοντέλου*, οι οποίοι εκτελούν μια επαναληπτική διαδικασία αναζήτησης ενός μοντέλου για την αναπαράσταση των δεδομένων. Για μία συγκεκριμένη αναπαράσταση μοντέλου εκτελείται η μέθοδος αναζήτησης παραμέτρων και εκτιμάται η ποιότητα του συγκεκριμένου μοντέλου.

1.5 Βασικές εργασίες εξόρυξης γνώσης

Οι μέθοδοι που χρησιμοποιούνται για την επίτευξη των στόχων του data mining εκτελούν κατά την εφαρμογή τους ένα σύνολο από εργασίες, οι βασικότερες εκ των οποίων είναι οι ακόλουθες:

- Ταξινόμηση *(Classification)*

- Τμηματοποίηση *(Clustering)*
- Εξαγωγή κανόνων συσχέτισης *(assosiation rules extraction)*
- Πρόβλεψη *(Prediction)*

Στη συνέχεια αναλύουμε κάθε επιμέρους εργασία και παραθέτουμε ενδεικτικά παραδείγματα για καλύτερη κατανόηση.

1.5.1 Ταξινόμηση *(Classification)*

Δοθέντων

- ενός προβλήματος με *N* κλάσεις: *C1, C2, ...Cn* όπου κάθε στιγμιότυπο του προβλήματος έχει *m* ιδιότητες (γνωρίσματα): *A1, A2,... Am*
- και ενός συνόλου στιγμιότυπων του προβλήματος για τα οποία γνωρίζουμε εκ των προτέρων σε ποια κλάση ανήκουν -το σύνολο αυτό είναι γνωστό ως

σύνολο εκπαιδευτικών στιγμιότυπων *(training set),*

το ζητούμενο είναι

- η δημιουργία ενός μοντέλου για την ταξινόμηση νέων άγνωστων στιγμιότυπων του προβλήματος. Με τον όρο ταξινόμηση εννοούμε την τοποθέτηση ενός στιγμιότυπου σε μία από τις προκαθορισμένες κλάσεις του προβλήματος.

Η επιτυχής έκβαση της ταξινόμησης εξαρτάται από δύο βασικούς παράγοντες:

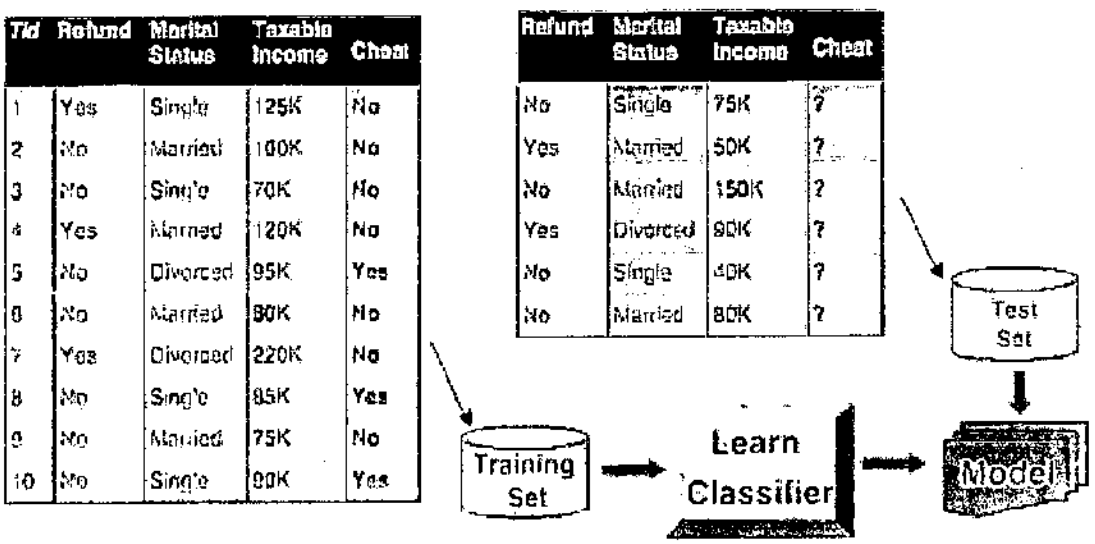
- *το σαφή καθορισμό των κλάσεων του προβλήματος* -οι κλάσεις είναι προκαθορισμένες και δεν μεταβάλλονται κατά τη διάρκεια της ταξινόμησης.
- *την «ποιότητα» του συνόλου των στιγμιότυπων εκπαίδευσης* – τα στιγμιότυπα αυτά θα πρέπει να είναι αντιπροσωπευτικά του προβλήματος.

Όπως έχουμε ήδη αναφέρει το σύνολο των εκπαιδευτικών στιγμιότυπων χρησιμοποιείται για την κατασκευή του ταξινομητή. Υπάρχει ωστόσο ένα ακόμη

σύνολο στιγμιότυπων, το **σύνολο των στιγμιότυπων ελέγχου** *(test set)* βάσει του οποίου ελέγχεται η απόδοση του ταξινομητή. Με τον όρο απόδοση εννοούμε την ακρίβεια με την οποία ο ταξινομητής απαντά στο πρόβλημα της ταξινόμησης νέων άγνωστων στιγμιότυπων του προβλήματος. Η απόδοση ισούται με τον αριθμό των

στιγμιότυπων του συνόλου ελέγχου για τα οποία ο ταξινομητής προέβλεψε σωστά την κλάση προς το συνολικό αριθμό των στιγμιότυπων του συνόλου ελέγχου.

Στο ακόλουθο σχήμα (Σχήμα 3) διαφαίνεται ο ρόλος των δύο επιμέρους συνόλων.



Σχήμα 3. Η λειτουργία της ταξινόμησης

Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα ταξινόμησης:

- **Παράδειγμα πολιτικής πίστωσης μιας τράπεζας**

Στην περίπτωση της πολιτικής πίστωσης, η τράπεζα θα ήθελε να γνωρίζει πότε μπορεί να δίνει δάνειο σε κάποιο πελάτη χωρίς να λαμβάνει μεγάλο ρίσκο.

Εφαρμόζοντας τη διαδικασία της ταξινόμησης το πρόβλημα ορίζεται ως εξής:

- *Οι πελάτες χωρίζονται στις κλάσεις: "άριστος", "καλός", "μέτριος" και "κακός" ανάλογα με την πιθανότητα κάθε πελάτη να εξοφλήσει το δάνειο.*
- *Κάθε πελάτης χαρακτηρίζεται από την ηλικία του, την εκπαίδευσή του, το ετήσιο*

εισόδημά του, κ.α.

- *Έχουμε στη διάθεση μας δεδομένα πελατών που έχουν δανειστεί από την τράπεζα στο παρελθόν.*

Ένα ενδεχόμενο αποτέλεσμα της ταξινόμησης για τις κλάσεις "άριστος" και "καλός" θα μπορούσε να είναι το ακόλουθο:

Για κάθε πελάτη P, με P.πτυχίο = μεταπτυχιακό and P.εισόδημα > 75,000 →P.κλάση = άριστος
Για κάθε πελάτη P, με P.πτυχίο = πτυχίο πανεπιστημίου ή (P.εισόδημα≥25,000 και P.εισόδημα ≤ 75000) →P.κλάση = καλός

Αξιοποιώντας το αποτέλεσμα μιας τέτοιας ταξινόμησης η τράπεζα αναμένεται να μειώσει το ρίσκο ο πελάτης στον οποίο χορήγησε κάποιο δάνειο να είναι ασυνεπής ως προς την εξόφληση του δανείου.

- **Παράδειγμα οργάνωσης διαφημιστικής καμπάνιας**

Στην περίπτωση της οργάνωσης μιας διαφημιστικής καμπάνιας, η εταιρία θα ήθελε να γνωρίζει ποιοι πελάτες είναι πιο πιθανό να απαντήσουν θετικά στην καμπάνια. Στόχος της εταιρίας είναι να προωθήσει την καμπάνια μόνο σε (πιθανά) ενδιαφερόμενα άτομα μειώνοντας έτσι το συνολικό κόστος.

Εφαρμόζοντας τη διαδικασία της ταξινόμησης το πρόβλημα ορίζεται ως εξής:

- *Οι πελάτες χωρίζονται στις κλάσεις: θετικοί και αρνητικοί αποδέκτες διαφημιστικών φυλλαδίων.*
- *Κάθε πελάτης χαρακτηρίζεται από το όνομά του, την ηλικία του, το επάγγελμά του κ.α.*
- *Έχουμε στη διάθεση μας δεδομένα πελατών που είχαν απαντήσει σε παλαιότερες διαφημιστικές καμπάνιες της εταιρίας.*

Ένα ενδεχόμενο αποτέλεσμα της ταξινόμησης για την κλάση των θετικών αποδεκτών διαφημιστικών φυλλαδίων θα μπορούσε να είναι το ακόλουθο:

Για κάθε πελάτη P, με (P.ηλικία>25 και P.ηλικία<55) και Περιοχή = Ν. Προάστεια

→P.κλάση = θετικός αποδέκτης

Η εταιρία θα μπορούσε να αξιοποιήσει το αποτέλεσμα αποστέλλοντας το νέο διαφημιστικό υλικό μόνο στους θετικούς αποδέκτες μειώνοντας έτσι το συνολικό κόστος της διαφημιστικής καμπάνιας.

1.5.2 Τμηματοποίηση *(Clustering)*

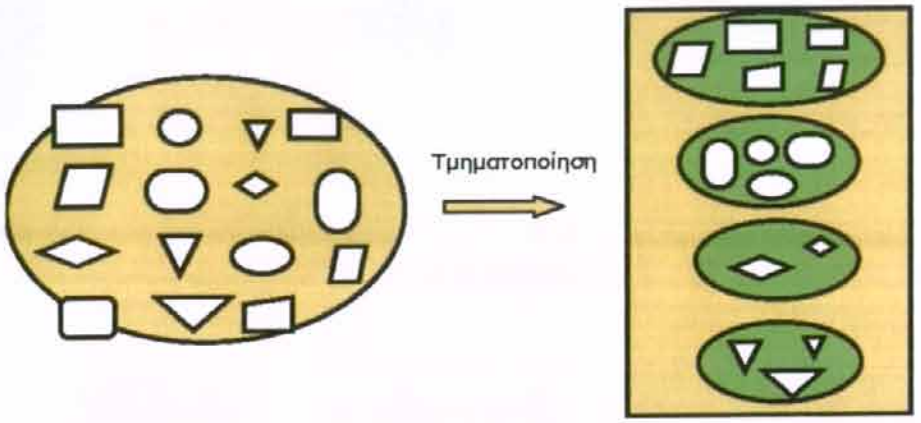
Δοθέντων

- ενός προβλήματος όπου κάθε στιγμιότυπο του προβλήματος έχει *m* ιδιότητες (γνωρίσματα): *A1, A2, ... Am*
- και ενός συνόλου στιγμιότυπων του προβλήματος το ζητούμενο είναι

- ο διαχωρισμός των στιγμιότυπων του προβλήματος σε τμήματα *(clusters)*, έτσι ώστε στιγμιότυπα με παρόμοια χαρακτηριστικά να ανήκουν στο ίδιο τμήμα και
- η εύρεση του προφίλ κάθε τμήματος.

Η τμηματοποίηση είναι κατάλληλη για την εύρεση τμημάτων αντικειμένων με παρόμοια χαρακτηριστικά. Έτσι όταν θέλουμε να εξάγουμε κανόνες σχετικά με τη συμπεριφορά των αντικειμένων ενός συγκεκριμένου τμήματος δε χρειάζεται να εξετάσουμε τις ανεξάρτητες εγγραφές του συνόλου των δεδομένων, αρκεί να εξετάσουμε τα χαρακτηριστικά του συγκεκριμένου τμήματος. Η ιδέα / διαίσθηση είναι πως τα στοιχεία που ανήκουν στο ίδιο τμήμα θα συμπεριφέρονται με ενιαίο τρόπο, καθώς έχουν παρόμοια χαρακτηριστικά. Συνεπώς, ένας κανόνας που είναι έγκυρος για κάποιο από τα στοιχεία ενός τμήματος αναμένεται να είναι έγκυρος και για τα υπόλοιπα στοιχεία του τμήματος.

Στο ακόλουθο σχήμα (Σχήμα 4) παρατίθεται ένα γενικό παράδειγμα τμηματοποίησης των ετερογενών στιγμιότυπων ενός προβλήματος σε τμήματα με κοινά χαρακτηριστικά.



Τα στιγμιότυπα του προβλήματος Τα εξαγόμενα τμήματα *(clusters)*

Σχήμα 4. Η λειτουργία της τμηματοποίησης

Η βασική διαφορά μεταξύ της ταξινόμησης και της τμηματοποίησης έγκειται στο γεγονός ότι στην ταξινόμηση οι κλάσεις είναι προκαθορισμένες, ενώ στην τμηματοποίηση δεν υπάρχουν προκαθορισμένες κλάσεις, τα στιγμιότυπα διασπόνται σε τμήματα βάσει της ομοιότητας που παρουσιάζουν μεταξύ τους ως προς τα γνωρίσματα της τμηματοποίησης. Συνεπώς, αν εφαρμόσουμε τμηματοποίηση σε ένα σύνολο δεδομένων, δεν υπάρχει κάποιο συγκεκριμένο σύνολο παραδειγμάτων το οποίο θα μπορούσε να μας υποδείξει ποιες είναι οι επιθυμητές σχέσεις που θα πρέπει να ισχύουν μεταξύ των δεδομένων.

Αρκετά συχνά η τμηματοποίηση χρησιμοποιείται και σαν πρώτο βήμα σε κάποια άλλη μορφή data mining εργασίας. Για παράδειγμα, μπορεί να χρησιμοποιηθεί σαν πρώτο βήμα στην προσπάθεια μερισμού της αγοράς. Αντί δηλαδή να προσπαθήσουμε να προσδιορίσουμε τι είδους διαφήμιση ταιριάζει καλύτερα σε κάθε πελάτη, μπορούμε να διασπάσουμε τους πελάτες σε τμήματα με βάση τις συνήθειές τους κατά την αγορά προϊόντων, να φτιάξουμε το προφίλ κάθε τμήματος και στη συνέχεια να προσδιορίσουμε το είδος της διαφήμισης που ταιριάζει καλύτερα στο κάθε τμήμα.

Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα τμηματοποίησης:

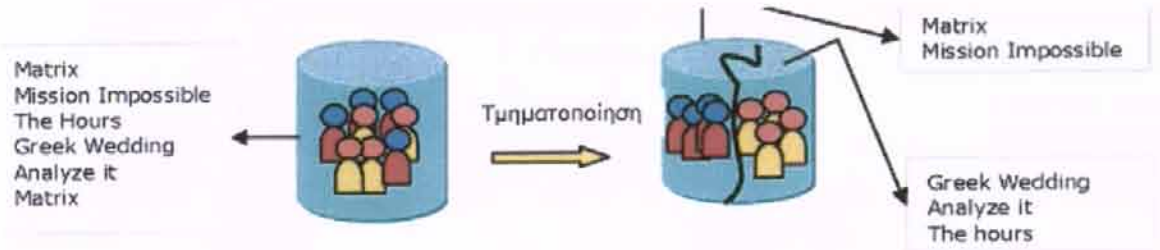
- **Παράδειγμα τμηματοποίησης του πληθυσμού μιας χώρας σε διάφορα μορφωτικά επίπεδα**

Στην περίπτωση αυτή κάθε στιγμιότυπο του προβλήματος έχει ιδιότητες όπως ηλικία,

τόπος κατοικίας, οικονομική κατάσταση, μόρφωση, κ.α., οπότε εφαρμόζοντας τμηματοποίηση μπορούμε να διασπάσουμε τον πληθυσμό με βάση τα γνωρίσματα αυτά και να βρούμε το προφίλ του κάθε τμήματος.

- **Παράδειγμα διαχωρισμού των χρηστών ενός δικτυακού τόπου με βάση τα κινηματογραφικά τους ενδιαφέροντα**

Στην περίπτωση αυτή κάθε στιγμιότυπο του προβλήματος έχει ιδιότητες όπως ηλικία, προηγούμενες προτιμήσεις σε ταινίες, επάγγελμα, μόρφωση, κ.α., οπότε μέσω της τμηματοποίησης μπορούμε να διασπάσουμε τους χρήστες σε τμήματα και να βρούμε



Σχήμα 5. Τμηματοποίηση των χρηστών με βάση τα κινηματογραφικά τους ενδιαφέροντα

Έτσι όταν εμφανίζεται ένας νέος χρήστης, μπορούμε να βρούμε το πιο κοντινό στο χρήστη τμήμα χρηστών (ανάλογα με τις προηγούμενες προτιμήσεις του) και να του προτείνουμε ταινίες με βάση τις προτιμήσεις του τμήματος στο οποίο ανήκει. Η διαίσθησή μας είναι πως θα τον ενδιαφέρουν ταινίες που ενδιαφέρουν και τα υπόλοιπα μέλη του τμήματος.

1.5.3 Εξαγωγή κανόνων συσχέτισης *(Association rules)*

Οι κανόνες συσχέτισης είναι κατάλληλοι για την εύρεση συσχετίσεων μεταξύ διαφορετικών αντικειμένων. Ένας κανόνας συσχέτισης μεταξύ δύο αντικειμένων A και B δηλώνει πως η παρουσία του A σε κάποιο στιγμιότυπο του προβλήματος συνεπάγεται και την παρουσία του B στο ίδιο στιγμιότυπο του προβλήματος και συμβολίζεται με $A \rightarrow B$.

Η εξαγωγή των κανόνων συσχέτισης γίνεται με τη βοήθεια κάποιων αλγορίθμων, οι οποίοι αποδεικνύονται αρκετά αποδοτικοί. Μετά την ανάλυση και την εύρεση των

κανόνων θα πρέπει να διαπιστωθεί κατά πόσο είναι έγκυροι και σημαντικοί για την εκάστοτε εφαρμογή. Για το σκοπό αυτό υπάρχουν δύο συντελεστές: η υποστήριξη *(support)* και η σιγουριά *(confidence)*.

- Η **υποστήριξη** *(support)* ισούται με το ποσοστό του συνόλου των στιγμιότυπων, έστω N το σύνολο των στιγμιότυπων, που ικανοποιεί το συνδυασμό Α και Β.

$$support = [AB]/N,$$

Έστω για παράδειγμα ο κανόνας συσχέτισης *γάλα* → *κατσαβίδια*, αν υποθέσουμε πως μόνο το 0.001 όλων των αγορών περιλαμβάνει γάλα και κατσαβίδια, τότε η υποστήριξη του κανόνα συσχέτισης είναι χαμηλή. Συνήθως, οι επιχειρήσεις δεν ενδιαφέρονται για κανόνες με χαμηλή υποστήριξη, δεδομένου ότι αφορούν ένα πολύ μικρό ποσοστό των πελατών τους.

Από την άλλη αν το 50% των αγορών περιλαμβάνει γάλα και ψωμί, τότε η υποστήριξη για τον κανόνα συσχέτισης *γάλα* → *ψωμί* είναι μεγάλη. Τέτοιοι κανόνες παρουσιάζουν ενδιαφέρον για τις επιχειρήσεις καθώς αφορούν ένα μεγάλο ποσοστό των πελατών.

- Η **σιγουριά** *(support)* ισούται με το ποσοστό του συνόλου των στιγμιότυπων για τα οποία όταν ισχύει το Α ισχύει και το Β.

$$confidence = [AB]/[A]$$

Για παράδειγμα, ο κανόνας συσχέτισης *ψωμί* → *γάλα* έχει μια σιγουριά 80% αν το 80% των αγορών που περιλαμβάνουν ψωμί περιλαμβάνει επίσης και γάλα. Για τις επιχειρήσεις ένας κανόνας με χαμηλή σιγουριά δεν παρουσιάζει ενδιαφέρον.

Να σημειώσουμε πως η σιγουριά του κανόνα *ψωμί* → *γάλα* μπορεί να διαφέρει από τη σιγουριά του κανόνα *γάλα* → *ψωμί*, παρόλο που και οι δύο κανόνες έχουν την ίδια υποστήριξη.

Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα κανόνων συσχέτισης:

- **Παράδειγμα: σχεδιασμός καταλόγου σε καταστήματα**

Τα μαγαζιά λιανικής πώλησης ενδιαφέρονται να βρουν συσχετίσεις μεταξύ των

διαφορετικών προϊόντων που αγοράζουν οι πελάτες του. Παραδείγματα τέτοιων συσχετίσεων θα μπορούσαν να είναι:

- Κάποιος που αγοράζει ψωμί είναι πολύ πιθανό να αγοράσει και γάλα (Σχήμα 6). Γνωρίζοντας ένα σουπερμάρκετ αυτόν τον κανόνα θα μπορούσε να τοποθετήσει σε διπλανά ράφια το ψωμί και το γάλα, δεδομένου ότι τα δύο αυτά προϊόντα αγοράζονται συχνά μαζί.

Σχήμα 6. Παράδειγμα του κανόνα συσχέτισης A → C που μπορεί να προκύψει από τις αγορές των πελατών ενός σουπερμάρκετ.

- Κάποιος που αγοράζει το βιβλίο "Database System Concepts" είναι πολύ πιθανό να αγοράσει και το βιβλίο "Operating System Concepts". Γνωρίζοντας ένα online βιβλιοπωλείο αυτόν τον κανόνα θα μπορούσε να προτείνει το βιβλίο "Operating System Concepts" στους πελάτες του που αγοράζουν το βιβλίο "Database System Concepts", δεδομένου ότι τα δύο αυτά βιβλία αγοράζονται συχνά μαζί.

Ένας κανόνας συσχέτισης πρέπει να σχετίζεται με κάποιο πληθυσμό: ο πληθυσμός αυτός αποτελείται από ένα σύνολο στιγμιότυπων. Στην περίπτωση του σουπερμάρκετ για παράδειγμα, ο πληθυσμός προκύπτει από το ιστορικό των αγορών των πελατών του και κάθε στιγμιότυπο του προβλήματος αποτελείται από τα προϊόντα που αγόρασε ο πελάτης κατά της διάρκεια μιας αγοράς. Στην περίπτωση του online καταστήματος από την άλλη, ο πληθυσμός αποτελείται από όλους τους πελάτες του καταστήματος και κάθε στιγμιότυπο του προβλήματος περιλαμβάνει τις προτιμήσεις και τα προϊόντα που αγόρασε ο πελάτης καθ' όλη τη διάρκεια λειτουργίας του καταστήματος.

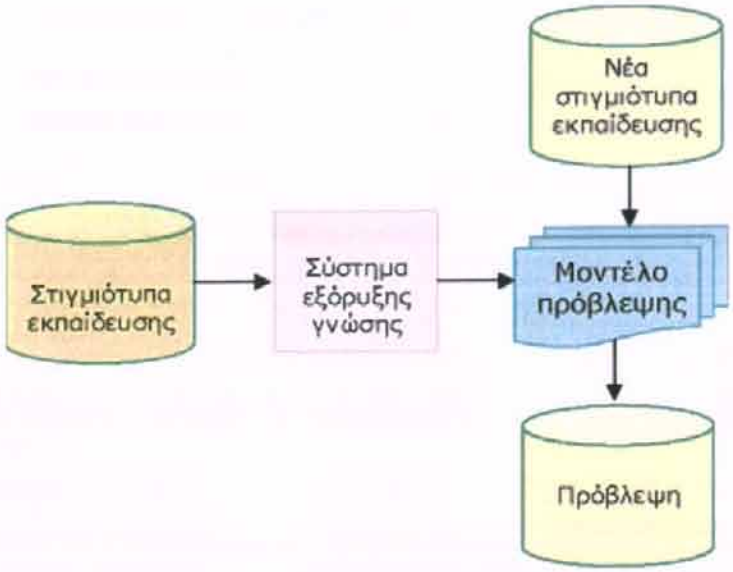
Παρατηρούμε λοιπόν ότι η έννοια του πληθυσμού καθορίζεται κάθε φορά από

το πρόβλημα που καλούμαστε να αντιμετωπίσουμε. Έτσι στο πρώτο παράδειγμα επικεντρωνόμαστε στις επιμέρους αγορές ενός πελάτη ενώ στο δεύτερο επικεντρωνόμαστε στη συνολική αγοραστική εικόνα του πελάτη χωρίς να λαμβάνουμε υπόψη τα προϊόντα που αγοράστηκαν στις επιμέρους αγορές.

1.5.4 Πρόβλεψη *(Prediction)*

Δοθέντος ενός μοντέλου πρόβλεψης και ενός νέου στιγμιότυπου του προβλήματος, το ζητούμενο είναι η πρόβλεψη της τιμής ενός συγκεκριμένου γνωρίσματος του στιγμιότυπου αυτού (Σχήμα 7).

Το μοντέλο πρόβλεψης «χτίζεται» μέσω των παραδειγμάτων του συνόλου εκπαίδευσης (πρόκειται για παραδείγματα στα οποία η τιμή του προς πρόβλεψη γνωρίσματος είναι γνωστή). Πέραν του συνόλου εκπαίδευσης υπάρχει και το σύνολο ελέγχου, το οποίο αποτελείται από παραδείγματα στα οποία η τιμή του προς πρόβλεψη γνωρίσματος είναι γνωστή – το σύνολο αυτό συνήθως ισούται αριθμητικά με το 1/3 των παραδειγμάτων του συνόλου εκπαίδευσης και χρησιμοποιείται για την αξιολόγηση του μοντέλου πρόβλεψης. Το σύνολο ελέγχου ελέγχει κατά κάποιο τρόπο την απόδοση του μοντέλου πρόβλεψης, δηλαδή την ακρίβεια με την οποία το μοντέλο πρόβλεψης προβλέπει την τιμή ενός άγνωστου γνωρίσματος στα νέα στιγμιότυπα του προβλήματος. Για να βρούμε την άγνωστη τιμή ενός γνωρίσματος σε κάποιο νέο στιγμιότυπο του προβλήματος, θα πρέπει να περάσουμε το στιγμιότυπο αυτό από το μοντέλο πρόβλεψης.



Σχήμα 7. Λειτουργία της πρόβλεψης

Συγκρίνοντας την πρόβλεψη με την ταξινόμηση που είδαμε στην αμέσως προηγούμενη ενότητα μπορούμε να πούμε πως η ταξινόμηση αποτελεί μια ειδική περίπτωση πρόβλεψης, καθώς αναφέρεται στην πρόβλεψη της κλάσης των στιγμιότυπων του προβλήματος.

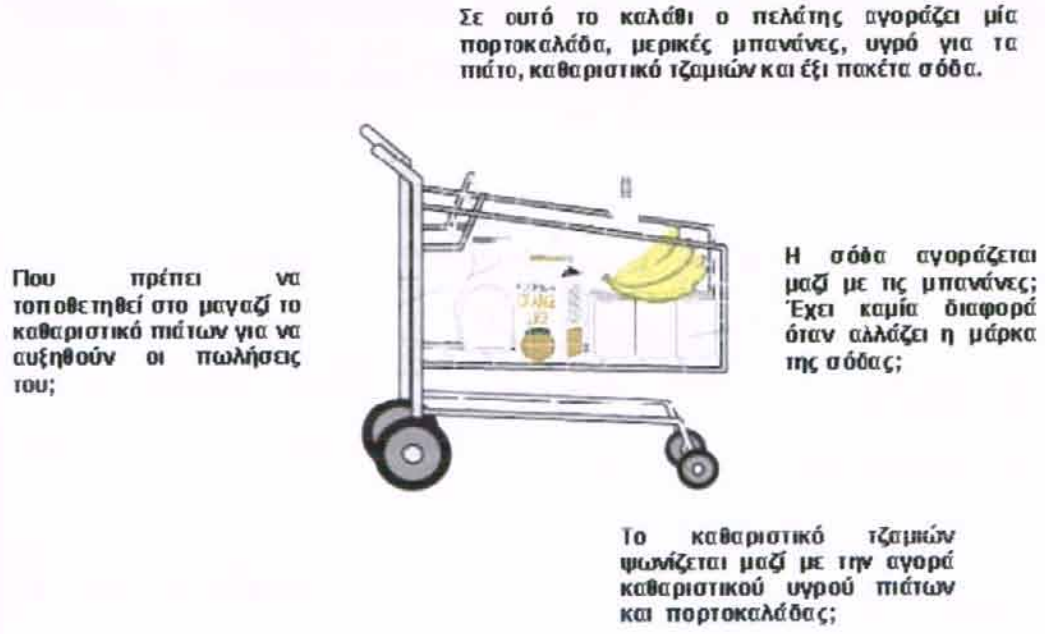
Μερικά ενδεικτικά παραδείγματα πρόβλεψης είναι: πρόβλεψε την πιθανότητα ένας ασθενής να πάσχει από μία συγκεκριμένη ασθένεια, πρόβλεψε το πλήθος των αγορών που θα κάνει ένας νέος πελάτης στον πρώτο χρόνο κ.α.

Πρέπει να τονιστεί ότι παρούσα πτυχιακή εργασία επικεντρώνεται στους αλγόριθμους εξαγωγής κανόνων συσχέτισης.

2. ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ ΚΑΙ Η ΑΝΑΛΥΣΗ ΚΑΛΑΘΙΟΥ ΤΗΣ ΝΟΙΚΟΚΥΡΑΣ

2.1 Εισαγωγή

Για να γίνει κατανοητή η έννοια του παραδείγματος της Ανάλυσης καλαθιού της Νοικοκυράς (*Market Basket Analysis*) θα χρησιμοποιηθεί το παράδειγμα του σχήματος 8. όπου ένας βιαστικός καταναλωτής αγόρασε κάποια σημαντικά προϊόντα για αυτόν από ένα super-market. Το καλάθι αυτό περιέχει κάποια προϊόντα που είναι ταξινομημένα – πορτοκαλάδα, μπανάνες, σόδα, καθαριστικό για τα τζάμια και απορρυπαντικό – και μας δείχνει τι αγόρασε ένας πελάτης σε μία επίσκεψή του στο συγκεκριμένο κατάστημα. Το ένα καλάθι λέει τι αγόρασε ένας πελάτης, αλλά όλες οι αγορές που έγιναν από όλους τους πελάτες περιέχουν σημαντική πληροφορία. Οι πελάτες δεν είναι απαραίτητο να είναι οι ίδιοι. Κάθε πελάτης αγοράζει διαφορετικά σύνολα προϊόντων, σε διαφορετικές ποσότητες, σε διαφορετικές χρονικές στιγμές κατά την διάρκεια μιας εβδομάδας. Η τεχνική της Ανάλυσης καλαθιού της Νοικοκυράς χρησιμοποιεί την πληροφορία για το τι αγόρασαν οι πελάτες με σκοπό να γίνει γνωστό «ποιοι είναι οι πελάτες» (ως προς τις συνήθειες) και γιατί κάνουν συγκεκριμένες αγορές.



Σχήμα 8. Παράδειγμα super-market

Η Ανάλυση καλαθιού της Νοικοκυράς δίνει την πληροφορία για το ποια προϊόντα τείνουν να αγοράζονται μαζί και ποια είναι τα προϊόντα που είναι αναγκαίο να διαφημιστούν. Αυτή η πληροφορία είναι πολύ σημαντική: Μπορεί να προτείνει την δημιουργία καινούργιων καταστημάτων σε περιοχές (συνοικίες κτλ.) όπως επίσης να δείξει ποια προϊόντα χρειάζονται ειδική συμπεριφορά (όπως προσφορές, κουπόνια κτλ.)

Παρόλο που η Ανάλυση καλαθιού της Νοικοκυράς έχει δημιουργηθεί για καταστήματα λιανικών πωλήσεων (super-market) έχει και αλλού εφαρμογές:

- Προϊόντα που αγοράζονται με πιστωτικές κάρτες, όπως ενοικίαση αυτοκινήτων ή δωματίων ενός ξενοδοχείου, βοηθούν στην πρόβλεψη για το ποιο προϊόν είναι πιθανόν να αγοραστεί από τον πελάτη σε επόμενες αγορές.
- Προαιρετικές υπηρεσίες που δίνονται από μία εταιρία τηλεπικοινωνιών σε έναν πελάτη (αναμονή κλήσεων, φραγή, εκτροπή κτλ.) μπορούν να δώσουν πακέτα προσφορών για να ανεβάσουν το εισόδημα της εταιρίας (π.χ. με την ενεργοποίηση της αναγνώρισης κλήσεων δώρο την φραγή κλήσεων).
- Τραπεζικές υπηρεσίες που χρησιμοποιούνται από πελάτες (δάνεια, υπηρεσίες επενδύσεων, τραπεζικοί λογαριασμοί) μπορούν να δώσουν την πληροφορία για το ποιες άλλες υπηρεσίες επιθυμεί ο κάθε πελάτης.
- Ιστορικά ασθενών σε νοσοκομεία αποκαλύπτουν περιπλοκές που βασίζονται σε συνδυασμούς θεραπευτικών αγωγών. (π.χ. έγκυες γυναίκες δεν μπορούν να λαμβάνουν οποιαδήποτε φάρμακα).

2.2 Πότε είναι η Ανάλυση καλαθιού της Νοικοκυράς χρήσιμη;

Η Ανάλυση καλαθιού της Νοικοκυράς διακρίνεται για την σαφήνεια και τη χρησιμότητα των αποτελεσμάτων της, τα οποία είναι υπό μορφή κανόνων συσχέτισης (*association rules)*. Υπάρχει μια προσφυγή σε έναν κανόνα συσχέτισης επειδή εκφράζει πώς τα προϊόντα και οι υπηρεσίες συσχετίζονται το ένα το άλλο, πώς τείνουν να ομαδοποιούνται. Ένας κανόνας όπως, «Εάν ένας πελάτης αγοράζει την υπηρεσία ταυτόχρονης ομιλίας 3 ατόμων (συνδιάσκεψης), κατόπιν εκείνος ο πελάτης

θα αγοράσει επίσης και την αναμονή κλήσης», είναι σαφής. Ακόμα καλύτερα, η Ανάλυση καλαθιού της Νοικοκυράς προτείνει ένα συγκεκριμένο σχέδιο δράσης, όπως την πρόσφορα της υπηρεσίας ταυτόχρονης ομιλίας 3 ατόμων και της αναμονής κλήσης σε ένα ενιαίο πακέτο υπηρεσιών.

Ενώ οι κανόνες συσχέτισης είναι εύκολο να γίνουν κατανοητοί, δεν είναι πάντα χρήσιμοι. Οι ακόλουθοι τρεις κανόνες είναι παραδείγματα πραγματικών κανόνων που παράγονται από τα παρακάτω πραγματικά στοιχεία:

- Τις Πέμπτες, οι καταναλωτές καταστημάτων supermarket αγοράζουν συχνά πάνες και μπύρα μαζί.
- Οι πελάτες που κάνουν συμφωνίες για υπηρεσίες συντήρησης είναι πολύ πιθανό να αγοράσουν μεγάλες συσκευές.
- Όταν ένα νέο κατάστημα που πουλάει διάφορα εργαλεία και είδη υγιεινής ανοίγει, ένα από τα συνηθέστερα προϊόντα που πωλούνται είναι καπάκια τουαλετών.

Αυτά τα τρία παραδείγματα επεξηγούν τους τρεις κοινούς τύπους κανόνων που παράγονται από την Ανάλυση καλαθιού της Νοικοκυράς: χρήσιμος κανόνας, τετριμμένος (trivial), και ανεξήγητος.

Ο χρήσιμος κανόνας περιέχει τις υψηλής ποιότητας πληροφορίες. Στην πραγματικότητα, μόλις βρεθεί το πρότυπο, είναι συχνά δύσκολο να δικαιολογηθεί. Ο κανόνας για τις πάνες και την μπύρα τις Πέμπτες προτείνει ότι την Πέμπτη το βράδυ, τα νεαρά ζευγάρια προετοιμάζονται για το Σαββατοκύριακο με το να εφοδιάσουν με πάνες για τα νήπια και μπύρα για τον μπαμπά (που, παραδείγματος χάριν, υποθέτουμε ότι παρακολουθεί ποδόσφαιρο την Κυριακή με τη συντροφιά μιας μπύρας). Ακόμα σημαντικότερο είναι το ότι οι διευθυντές μπορούν τώρα να λάβουν μέτρα. με το να τοποθετήσουν τη δικιά τους μάρκα πανών τους κοντά στο διάδρομο που περιέχει την μπύρα, μπορούν να αυξήσουν τις πωλήσεις ενός προϊόντος. Επειδή ο κανόνας γίνεται κατανοητός εύκολα, προτείνει τις εύλογες αιτίες, που οδηγούν σε άλλες επεμβάσεις: τοποθετώντας άλλα μωρουδιακά προϊόντα δίπλα στις μπύρες υπενθυμίζει έτσι τους πελάτες τι άλλο πρέπει να αγοράσουν μαζί με τη μπύρα βάζοντας τρόφιμα, όπως τσιπς πατάτας κοντά στα προϊόντα μωρών.

Τα τετριμμένα αποτελέσματα είναι γνωστά ήδη από κάποιον εξοικειωμένο με τον κόσμο των επιχειρήσεων. Το δεύτερο παράδειγμα Οι πελάτες που κάνουν συμφωνίες για υπηρεσίες συντήρησης είναι πολύ πιθανό να αγοράσουν μεγάλες συσκευές.) είναι ένα παράδειγμα ενός τετριμμένου κανόνα. Στην πραγματικότητα, ήδη ξέρουμε ότι οι πελάτες αγοράζουν τις συμβόλαια συντήρησης και τις μεγάλες συσκευές συγχρόνως. Για πιο λόγο άλλωστε να αγόραζαν τις συμβόλαια συντήρησης; Τα συμβόλαια συντήρησης διαφημίζονται μαζί με τις μεγάλες συσκευές και πωλούνται σπάνια χωριστά. Αυτός ο κανόνας, εν τούτοις, βασίστηκε στην ανάλυση πραγματικών στοιχείων εκατοντάδων χιλιάδων συναλλαγών. Αν και είναι έγκυρο και βασισμένο σε πραγματικά δεδομένα, είναι άχρηστο. Παρόμοια αποτελέσματα αφθονούν: Πελάτες που αγοράζουν χρώμα αγοράζουν και βούρτσες ή πινέλα, το πετρέλαιο και τα φίλτρα πετρελαίου αγοράζονται μαζί όπως είναι και τα χάμπουργκερ μαζί με τα ψωμάκια χάμπουργκερ.

Ένα λεπτότερο πρόβλημα εμπίπτει στην ίδια κατηγορία. Ένα φαινομενικά ενδιαφέρον αποτέλεσμα, όπως το γεγονός ότι οι άνθρωποι που αγοράζουν την υπηρεσία ταυτόχρονης ομιλίας 3 ατόμων, κατόπιν εκείνοι οι πελάτες θα αγοράσουν επίσης και την αναμονή κλήσης μπορεί να είναι το αποτέλεσμα των προγραμμάτων μάρκετινγκ και των πακέτων πρόσφορων. Στην περίπτωση των επιλογών τηλεφωνικών υπηρεσιών, η υπηρεσία ταυτόχρονης ομιλίας 3 ατόμων ομαδοποιείται χαρακτηριστικά με την αναμονή κλήσης, και έτσι είναι δύσκολο να παραγγελθεί ξεχωριστά. Σε αυτήν την περίπτωση, η ανάλυση δεν παράγει τα αξιόλογα αποτελέσματα παράγει, αποτελέσματα γνωστά από πριν που έχουν ήδη εκμεταλλευτεί. Αν και είναι κίνδυνος για οποιαδήποτε τεχνική Εξόρυξης Δεδομένων, η Ανάλυση καλαθιού της Νοικοκυράς είναι ιδιαίτερα ευαίσθητη στο να αναπαράγει αποτελέσματα προηγούμενων διαφημιστικών εκστρατειών λόγω της εξάρτησής της από δεδομένα πωλήσεων, αυτά ακριβώς τα δεδομένα που καθορίζουν την επιτυχία της εκστρατείας. Τα αποτελέσματα από την Ανάλυση καλαθιού της Νοικοκυράς μπορούν απλά να «αναμασούν» την επιτυχία των προηγούμενων εκστρατειών μάρκετινγκ.

Ανεξήγητα αποτελέσματα φαίνονται να μην έχουν καμία εξήγηση και δεν προτείνουν κανένα σχέδιο δράσης. Το τρίτο πρότυπο (Όταν ένα νέο κατάστημα που πουλάει εργαλεία και είδη υγιεινής ανοίγει, ένα από τα συνηθέστερα προϊόντα που

πωλούνται είναι καπάκια τουαλετών) μας ενημερώνει για ένα γεγονός αλλά παρέχει πληροφορίες που δεν προσφέρουν διορατικότητα στην κατανόηση της συμπεριφοράς των καταναλωτών ή των εμπορευμάτων, και δεν προτείνουν περαιτέρω ενέργειες. Σε αυτήν την περίπτωση, μια μεγάλη εταιρία μπορεί να ανακαλύψει ένα τέτοιο πρότυπο, αλλά μπορεί να μη καταφέρει να υπολογίσει πώς να ωφεληθεί από αυτό. Πολλά προϊόντα πωλούνται με έκπτωση κατά τη διάρκεια των εγκαινίων του καταστήματος, αλλά τα καπάκια τουαλετών ξεχωρίζουν. Περισσότερη έρευνα μπορεί να δώσει κάποια εξήγηση: Είναι η έκπτωση στα καπάκια τουαλετών πολύ μεγαλύτερη απ' ό,τι για άλλα προϊόντα; Τοποθετούνται με συνέπεια σε μια περιοχή υψηλής κυκλοφορίας για τα εγκαίνια του καταστήματος αλλά κρύβονται άλλες μέρες; Είναι δύσκολο να βρεθούν άλλες μέρες; Οποιαδήποτε και να 'ναι η αιτία, είναι αμφισβητήσιμο αν η περαιτέρω ανάλυση ακριβώς των δεδομένων της αγοράς μπορεί να δώσει μια αξιόπιστη εξήγηση.

Κατά την εφαρμογή της Ανάλυσης καλαθιού της Νοικοκυράς, πολλά από τα αποτελέσματα είναι είτε συχνά είτε τετριμμένα είτε ανεξήγητα. Οι τετριμμένοι κανόνες αναπαράγουν κοινή γνώση για την επιχείρηση, που σπαταλά προσπάθεια και δύναμη που θα μπορούσε να χρησιμοποιηθεί για να εφαρμόσει περισσότερο εξεζητημένες και περίπλοκες τεχνικές ανάλυσης. Συχνά, τα τετριμμένα αποτελέσματα αναπαράγουν απλά τα προηγούμενα αποτελέσματα, όπως στις εκστρατείες μάρκετινγκ, αλλά δεν παρέχουν καμία οδηγία για τις μελλοντικές ενέργειες. Οι ανεξήγητοι κανόνες δεν είναι χρήσιμοι. Μπορεί να χρειαστούν περαιτέρω έρευνα έξω από τη σφαίρα της εξόρυξης δεδομένων για να τους καταλάβει κανείς καλύτερα. Ο καθορισμός του ποιοι κανόνες είναι πολύτιμοι μπορεί να απαιτήσει την εκμετάλλευση γνώση προηγούμενων διαφημιστικών εκστρατειών, ομαδοποίηση των υπηρεσιών, καθώς και άλλων εξωτερικών ή και ιστορικών παραγόντων.

Η Ανάλυση Καλαθιού της Νοικοκυράς χρησιμοποιείται συνήθως για να κάνει συγκρίσεις μεταξύ των θέσεων μέσα σε μια ενιαία αλυσίδα. Ο κανόνας για τις πωλήσεις καπακιων τουαλετών στα καταστήματα ειδών υγιεινής είναι ένα παράδειγμα όπου οι πωλήσεις στα νέα καταστήματα συγκρίνονται με τις πωλήσεις στα υπάρχοντα καταστήματα. Διαφορετικά καταστήματα εκθέτουν διαφορετικά πρότυπα πωλήσεων για πολλούς λόγους: περιφερειακές τάσεις, αποτελεσματικότητα

της διαχείρισης, και δημογραφικά πρότυπα στην περιοχή μελέτης, παραδείγματος χάριν. Τα κλιματιστικά μηχανήματα και οι ανεμιστήρες αγοράζονται συχνά κατά τη διάρκεια των καυσώνων, αλλά τα κύματα καύσωνα έχουν επιπτώσεις μόνο σε μια περιορισμένη περιοχή.

Μέσα σε μικρότερες περιοχές όπου δημογραφικές μελέτες της περιοχής μελέτης θα μπορούσαν να ασκήσουν μεγάλη επίδραση, θα αναμέναμε τα καταστήματα στα πλούσια προάστια να εκθέσουν διαφορετικά πρότυπα πωλήσεων από εκείνα στις γειτονιές των πόλεων. Αυτά είναι παραδείγματα όπου η Ανάλυση Καλαθιού της Νοικοκυράς μπορεί να βοηθήσει στη περιγραφή των διαφορών και να χρησιμεύσει ως ένα παράδειγμα για την κατευθυνόμενη εξόρυξη δεδομένων (directed data mining).

Πώς χρησιμοποιείται η Ανάλυση Καλαθιού της Νοικοκυράς για να γίνουν αυτές οι συγκρίσεις; Κατ' αρχάς, πρέπει να αυξήσουμε τις συναλλαγές (transactions) στα δεδομένα με εικονικά στοιχεία (virtual items) που διευκρινίζουν από ποια ομάδα, όπως μια υπάρχουσα θέση ή μια νέα θέση, η συναλλαγή προέρχεται. Η βοήθεια των εικονικών αυτών στοιχείων περιγράφει τη συναλλαγή, αν και το εικονικό στοιχείο δεν είναι ένα προϊόν ή υπηρεσία. Παραδείγματος χάριν, μια σειρά πωλήσεων σε ένα υπάρχον κατάστημα εξοπλισμού μπορεί να περιλαμβάνει τα ακόλουθα προϊόντα:

- Ένα σφυρί
- Ένα κιβώτιο με καρφιά
- Λεπτό γυαλόχαρτο

Αφού αυξήσουμε τα δεδομένα για να διευκρινίσουμε από που προήλθαν, η συναλλαγή γίνεται:

- Ένα σφυρί
- Ένα κιβώτιο με καρφιά
- Λεπτό γυαλόχαρτο
- "στο υπάρχον κατάστημα εξοπλισμού."

Αργότερα σε αυτό το κεφάλαιο, θα μιλήσουμε περισσότερο για τα εικονικά στοιχεία, πότε είναι χρήσιμα, και μερικές προειδοποιήσεις όσο αναφορά τη χρήση τους.

Για να συγκρίνει κάποιος τις πωλήσεις των νέα καταστήματα με αυτές των υπαρχόντων καταστημάτων, η διαδικασία είναι η εξής:

1. Συγκέντρωση δεδομένων για μια συγκεκριμένη περίοδο (όπως δύο εβδομάδες) από τα εγκαίνια των νέων καταστημάτων. Αύξηση κάθε μιας από τις συναλλαγές αυτών των δεδομένων με ένα εικονικό στοιχείο που λέει ότι η συναλλαγή είναι από ένα νέο κατάστημα.
2. Συγκέντρωση σχεδόν ίδιου ποσού δεδομένων από υπάρχοντα καταστήματα. Εδώ μπορεί να χρησιμοποιηθεί ένα δείγμα σε όλα τα υπάρχοντα καταστήματα ή να γίνουν δειγματοληπτικοί έλεγχοι σε όλα τα καταστήματα σε συγκρίσιμες θέσεις. Αύξηση των συναλλαγών σε αυτά τα δεδομένα με ένα εικονικό στοιχείο λέγοντας ότι η συναλλαγή είναι από ένα υπάρχον κατάστημα.
3. Εφαρμογή της τεχνικής της Ανάλυσης Καλαθιού της Νοικοκυράς ανάλυσης για να βρεθούν οι κανόνες συσχέτισης σε κάθε μια ομάδα.
4. Δίνουμε ιδιαίτερη προσοχή στους κανόνες συσχέτισης που περιέχουν τα εικονικά στοιχεία.

Οι κανόνες που παράγονται από την Ανάλυση Καλαθιού της Νοικοκυράς μπορούν να χαρακτηριστούν ως αφετηρίες για την περαιτέρω δόκιμες υποθέσεων. Γιατί ένα συγκεκριμένο πρότυπο υπάρχει στα υπάρχοντα καταστήματα και ένα άλλο στα νέα καταστήματα; Ο κανόνας για τα καπάκια τουαλετών και τα εγκαίνια καταστημάτων, παραδείγματος χάριν, προτείνει την εξέταση περισσότερο των πωλήσεων καπακιων τουαλετών στα υπάρχοντα καταστήματα σε διαφορετικές χρονικές στιγμές κατά τη διάρκεια του έτους.

2.3 Λειτουργία της μεθόδου

Η Ανάλυση Καλαθιού της Νοικοκυράς ξεκινά τη λειτουργία της με τις συναλλαγές που περιέχουν μια ή περισσότερες προσφορές προϊόντων ή υπηρεσιών και κάποιες στοιχειώδεις πληροφορίες για τη συναλλαγή. Για σκοπό της ανάλυσης, καλούμε τις προσφορές προϊόντων ή υπηρεσιών αντικείμενα. Ο παρακάτω πίνακας 9. επεξηγεί πέντε συναλλαγές σε ένα παντοπωλείο που έχει πέντε προϊόντα.

Πίνακας 9. Συναλλαγές για αγορές προϊόντων

Πελάτης	Αντικείμενα
1	χυμός από πορτοκάλι, σόδα
2	γάλα, χυμός από πορτοκάλι, καθαριστικό παραθύρων
3	χυμός από πορτοκάλι, απορρυπαντικό
4	χυμός από πορτοκάλι, απορρυπαντικό, σόδα
5	καθαριστικό παραθύρων, σόδα

Αυτές οι συναλλαγές απλοποιούνται για να περιλάβουν μόνο τα αντικείμενα που αγοράζονται. Πώς θα αξιοποιηθούν πληροφορίες όπως ημερομηνία και ο χρόνος που έγινε η αγορά και εάν ο πελάτης πλήρωσε με μετρητά θα συζητηθούν αργότερα σε αυτό το κεφάλαιο.

Κάθε μια από αυτές τις συναλλαγές μας δίνει τις πληροφορίες για το ποια προϊόντα αγοράσθηκαν με ποια άλλα προϊόντα. Χρησιμοποιώντας αυτά τα δεδομένα, μπορούμε να δημιουργήσουμε έναν πίνακα περιστατικών που συνέβησαν ταυτόχρονα που λέει πόσες φορές ένα οποιοδήποτε ζευγάρι προϊόντων αγοράστηκε μαζί

Πίνακας 10. Ταυτόχρονη αγορά προϊόντων

	Χυμός από πορτοκάλι	Καθαριστικό παραθύρων	Γάλα	Σόδα	Απορρυπαντικό
Χυμός από πορτοκάλι	4	1	1	2	1
Καθαριστικό παραθύρων	1	2	1	1	0
Γάλα	1	1	1	0	0
Σόδα	2	1	0	3	1
Απορρυπαντικό	1	0	0	1	2

Αυτός ο πίνακας 10 μας λέει τον αριθμό των φορών που δύο προϊόντα εμφανίζονται ταυτόχρονα σε μια συναλλαγή. Παραδείγματος χάριν, εξετάζοντας το κελί όπου η σειρά "Σόδα" τέμνει τη στήλη " χυμός από πορτοκάλι ", βλέπουμε ότι δύο συναλλαγές περιέχουν και τη σόδα και το χυμό από πορτοκάλι. Αυτό ελέγχεται

εύκολα στα αρχικά δεδομένα της συναλλαγής, όπου οι πελάτες 1 και 4 αγόρασαν και τα δύο αυτά αντικείμενα. Οι τιμές κατά μήκος της διαγώνιου (παραδείγματος χάριν, η αξία στην " χυμός από πορτοκάλι " στήλη και στην " χυμός από πορτοκάλι " σειρά) αντιπροσωπεύουν τον αριθμό συναλλαγών που περιέχουν ακριβώς εκείνο το αντικείμενο.

Ο πίνακας ταυτόχρονων αγορών περιέχει μερικά απλά πρότυπα σχέδια:

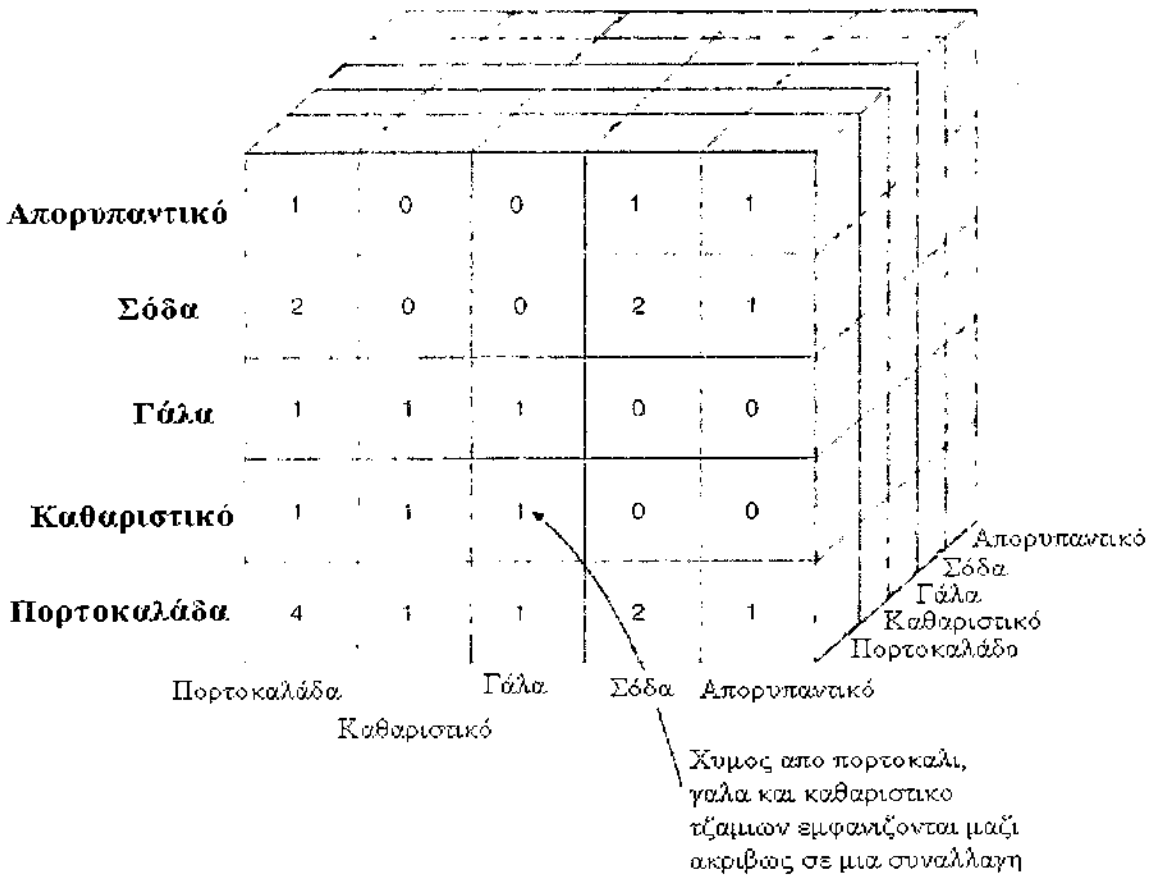
- Ο χυμός πορτοκαλιού και η σόδα είναι πιθανότερο να αγοραστούν μαζί από οποιαδήποτε άλλα δύο στοιχεία.
- Το απορρυπαντικό δεν αγοράζεται ποτέ με τον καθαριστικό παραθύρων ή το γάλα.
- Το γάλα δεν αγοράζεται ποτέ με τη σόδα ή το απορρυπαντικό.

Αυτές οι απλές παρατηρήσεις είναι παραδείγματα συσχετίσεων και μπορούν να προτείνουν έναν επίσημο κανόνα όπως: "Αν ένας πελάτης αγοράζει σόδα, ο πελάτης αγοράζει επίσης και γάλα." Για τώρα, θα αναβάλλουμε τη συζήτηση για το πώς βρίσκουμε αυτόν τον κανόνα αυτόματα. Άντ' αυτού, υποβάλλουμε την ερώτηση: Πόσο «καλός» είναι αυτός ο κανόνας; Στα δεδομένα, δύο από τις πέντε συναλλαγές περιλαμβάνουν και τη σόδα και το χυμό από πορτοκάλι. Αυτές οι δύο συναλλαγές λέμε ότι *υποστηρίζουν* τον κανόνα. Ένας άλλος τρόπος να εκφράσουμε αυτό είναι ως ποσοστό. Η υποστήριξη για τον κανόνα είναι δύο από πέντε ή 40 τοις εκατό.

Δεδομένου ότι και οι δύο συναλλαγές που περιέχουν τη σόδα περιέχουν επίσης και το χυμό από πορτοκάλι, υπάρχει ένας υψηλός βαθμός *εμπιστοσύνης* στον κανόνα. Στην πραγματικότητα, κάθε συναλλαγή που περιέχει τη σόδα περιέχει επίσης το χυμό από πορτοκάλι, έτσι ο κανόνας "εάν σόδα, τότε χυμός από πορτοκάλι" έχει μια εμπιστοσύνη 100 τοις εκατό. Είμαστε λιγότερο βέβαιοι για τον αντίστροφο κανόνα, "εάν χυμός από πορτοκάλι, τότε σόδα " λόγω του ότι από τις 4 συναλλαγές με το χυμό από πορτοκάλι, μόνο 2 εμπεριέχουν τη σόδα. Η εμπιστοσύνη της, τότε, είναι ακριβώς 50 τοις εκατό. Τυπικότερα, η εμπιστοσύνη είναι η αναλογία του αριθμού των συναλλαγών που υποστηρίζουν τον κανόνα προς τον αριθμό συναλλαγών όπου το υπό όρους μέρος του κανόνα κρατά. Ένας άλλος ορισμός είναι ότι η εμπιστοσύνη είναι η αναλογία του αριθμού συναλλαγών με όλα τα αντικείμενα

προς τον αριθμό συναλλαγών που περιχέουν μόνο τα "εάν" αντικείμενα.

Η ιδέα πίσω από τον πίνακα ταυτόχρονης εμφάνισης αντικειμένων μπορεί να επεκταθεί και για περισσότερα αντικείμενα, όχι μόνο για ζευγάρια. Για συνδυασμούς 3 αντικειμένων, ας υποθέσουμε ένα κύβο με τη κάθε πλευρά του να διαιρείται σε 5 διαφορετικά κομμάτια όπως φαίνεται και στο σχήμα 11. Οι συνδυασμοί που μπορούν να γίνουν είναι παρά πολλοί. Ας σκεφτούμε ότι και με 5 μέρη, υπάρχουν ήδη 125 διαφορετικοί υπό-κύβοι να γεμίσουν. Μεγαλύτερη ανάλυση εδώ ξεφεύγει από τους σκοπούς αυτής της εργασίας.



Σχήμα 11. Ταυτόχρονη εμφάνιση αντικειμένων

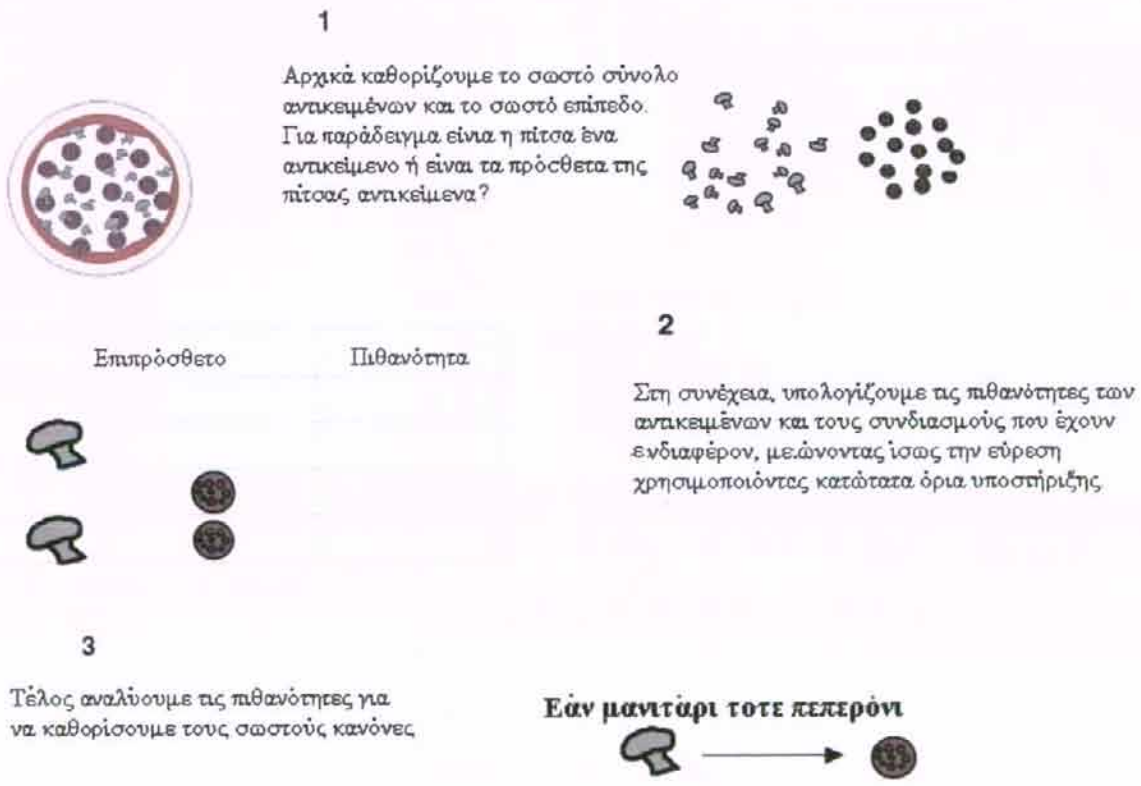
2.4 Η βασική διαδικασία

Επισκόπηση

Η βασική διαδικασία για την μέθοδο με τη βοήθεια ενός παραδείγματος από ένα κατάστημα που φτιάχνει πίτσες περιγράφεται στο σχήμα 12. Υπάρχουν τρεις

σημαντικές ανησυχίες που ανακύπτουν κατά τη χρήση της μεθόδου:

- Διαλέγοντας το σωστό σύνολο στοιχείων.
- Παράγοντας τους κανόνες με την ερμηνεία των αριθμήσεων στη μήτρα των ταυτόχρονων εμφανίσεων.
- Ξεπερνώντας τα πρακτικά όρια που επιβάλλονται από τις χιλιάδες ή τις δεκάδες χιλιάδες αντικειμένων που εμφανίζονται σε συνδυασμούς αρκετά πολλούς έτσι ώστε να είναι ενδιαφέροντα για περαιτέρω ανάλυση.



Σχήμα 12. Τα βασικά βήματα για την Ανάλυση Καλαθιού της Νοικοκυράς.

Αυτά τα 3 θέματα θα συζητηθούν παρακάτω.

Διαλέγοντας το σωστό στοιχειοσύνολο.

Τα δεδομένα που χρησιμοποιούνται για την μέθοδο είναι σε γενικές γραμμές τα λεπτομερή δεδομένα των συναλλαγών στα σημεία πώλησης. Η συγκέντρωση και η χρησιμοποίηση αυτών των δεδομένων είναι ένα κρίσιμο μέρος της εφαρμογής της μεθόδου, που εξαρτάται σημαντικά από τα αντικείμενα που επιλέγονται για την ανάλυση. Το τι αποτελεί ένα ιδιαίτερο στοιχείο εξαρτάται από την εκάστοτε επιχειρησιακή ανάγκη. Μέσα σε ένα παντοπωλείο όπου υπάρχουν δεκάδες χιλιάδων προϊόντα στα ράφια, μια παγωμένη πίτσα να θεωρηθεί στοιχείο για την ανάλυση ασχέτων των επιπρόσθετων της (extra τυρί, πεπερονι, ή μανιτάρια), της ζύμης της (extra παχιά, ολικής άλεσης, ή λευκό), ή του μεγέθους της. Έτσι, η αγορά μιας μεγάλης ολικής αλέσεως με χορταρικά πίτσας περιέχει το ίδιο "κατεψυγμένη πίτσα" αντικείμενο όπως και η αγορά μιας πεπερονι με πρόσθετο τυρί. Ένα δείγμα τέτοιων συναλλαγών φαίνεται στον πίνακα 13.

Πίνακας 13. Συναλλαγές με αθροισμένα αντικείμενα

Πελάτης	Πίτσα	Γάλα	Ζάχαρη	Μήλα	Καφές
1	√				
2		√	√		
3	√			√	√
4		√			√
5	√		√	√	√

Από την άλλη ο διευθυντής του τμήματος κατεψυγμένων φαγητών ή μιας αλυσίδας εστιατόριων μπορεί να ενδιαφέρεται πολύ για κάποιους συγκεκριμένους συνδυασμούς επιπρόσθετων που παραγγέλνονται μαζί με την αγορά μιας πίτσας. Μπορεί να αποσυνθέσει μια παραγγελιά για πίτσα όπως φαίνεται στον παρακάτω πίνακα.

Σε κάποιο μεταγενέστερο χρονικό σημείο, το εστιατόριο μπορεί να ενδιαφερθεί για τις συναλλαγές του με μεγαλύτερη λεπτομέρεια, έτσι το ενιαίο "κατεψυγμένη πίτσα" αντικείμενο δεν θα ήταν πλέον ικανοποιητικό. Ή, τα εστιατόρια να διευρύνουν τις επιλογές επιλογών τους και να γίνουν ενδιαφέρονται λιγότερο σε όλα τα διαφορετικά επιπρόσθετα. Τα αντικείμενα του ενδιαφέροντος

μπορούν να αλλάξουν κατά τη διάρκεια του χρόνου. Αυτό μπορεί να δημιουργήσει προβλήματα κατά την προσπάθεια να χρησιμοποιηθούν ιστορικά στοιχεία εάν τα δεδομένα της συναλλαγής έχουν συνοψιστεί. Η επιλογή του σωστού επιπέδου λεπτομέρειας είναι μια κρίσιμη εκτίμηση για την ανάλυση.

Πίνακας 14. Συναλλαγές με λεπτομερέστερα δεομένα

Πελάτης	Extra τυρί	Κρεμμύδια	Πιπεριές	Μανιτάρια	Ελιές
1	√	√			√
2			√		
3	√	√		√	
4		√			√
5	√		√	√	√

Οι ταξονομίες βοηθούν στη γενίκευση των αντικειμένων.

Σε πραγματικές συνθήκες, τα δεδομένα έχουν κωδικούς προϊόντων που εμπίπτουν σε ιεραρχικές κατηγορίες, αποκαλούμενο ταξονομίες (*taxonomies)*. Προσεγγίζοντας ένα πρόβλημα με την μέθοδο αυτή, ποιο επίπεδο της ταξονομίας είναι το σωστό να χρησιμοποιήσουμε;

Αυτό θέτει ζητήματα όπως:

- Είναι οι μεγάλες τηγανητές πατάτες και οι μικρές το ίδιο προϊόν;
- Είναι το εμπορικό σήμα του παγωτού πιο σχετικό από τη γεύση του;
- Ποιο είναι σημαντικότερο: το μέγεθος, το ύφος, το σχέδιο, ή ο σχεδιαστής ενός ρούχου;
- Είναι η επιλογή εξοικονόμησης ενέργειας σε μια μεγάλη συσκευή ενδεικτική της συμπεριφοράς πελατών;

Ο αριθμός συνδυασμών που μπορούμε να σκεφτούμε αυξάνεται πολύ γρήγορα καθώς ο αριθμός αντικειμένων που χρησιμοποιούνται στην ανάλυση αυξάνεται. Αυτό προτείνει τη χρήση αντικειμένων από τα πιο υψηλά επίπεδα της ταξονομίας, π.χ. "παγωμένα επιδόρπια" αντί "παγωτό". Από την άλλη, όσο πιο

συγκεκριμένα τα αντικείμενα είναι, τόσο πιθανότερο είναι τα αποτελέσματα να είναι χρήσιμα. Η γνώση του τι πωλείται μαζί με μια ιδιαίτερη μάρκα κατεψυγμένης πίτσας, παραδείγματος χάριν, μπορεί να βοηθήσει στη διαχείριση της σχέσης με τον παραγωγό. Ένας συμβιβασμός είναι να χρησιμοποιηθούν περισσότερο γενικά στοιχεία αρχικά, κατόπιν να επαναληφθεί η παραγωγή του κανόνα για να δοθεί περισσότερο προσοχή σε πιο συγκεκριμένα στοιχεία. δεδομένου ότι η ανάλυση εστιάζει στα πιο συγκεκριμένα στοιχεία, χρησιμοποιούμε μόνο το υποσύνολο των συναλλαγών που περιέχουν εκείνα τα στοιχεία.

Η πολυπλοκότητα ενός κανόνα αναφέρεται στον αριθμό αντικειμένων που περιέχει. Όσο περισσότερα αντικείμενα υπάρχουν στις συναλλαγές, τόσο πιο πολύ χρόνο χρειάζονται για να παραγάγουν κανόνες δεδομένης πολυπλοκότητας. Έτσι, η επιθυμητή πολυπλοκότητα των κανόνων καθορίζει επίσης πόσο συγκεκριμένα ή γενικευμένα τα αντικείμενα πρέπει να είναι.

Ποιότητα δεδομένων

Τα δεδομένα που χρησιμοποιούνται για την ανάλυση καλαθιού νοικοκυράς δεν είναι γενικά πολύ υψηλής ποιότητας. Συλλέγονται στο σημείο που γίνονται οι αγορές και χρησιμοποιούνται κυρίως για λειτουργικούς λόγους όπως τον έλεγχο καταλόγων. Τα δεδομένα από τα λειτουργικά συστήματα είναι συχνά «βρώμικα» και χρειάζονται εκτενές ξεκαθάρισμα πριν να γίνουν μια καλή πηγή για την εφαρμογή των αλγόριθμων εξόρυξης δεδομένων και υποστήριξη απόφασης. Τα δεδομένα είναι πιθανό να έχουν πολλαπλές εκδόσεις, διορθώσεις, ασυμβίβαστους τύπους κώδικα, και τα λοιπά. Τα διαφορετικά καταστήματα μέσα σε μια ενιαία αλυσίδα καταστημάτων έχουν μερικές φορές ελαφρώς διαφορετικές ιεραρχίες προϊόντων ή διαφορετικούς τρόπους αντιμετώπισης καταστάσεων όπως οι εκπτώσεις για παράδειγμα. Μερικά συστήματα είναι πιο ενημερωμένα από άλλα συστήματα. Αυτά τα προβλήματα είναι χαρακτηριστικά κατά χρησιμοποίηση οποιουδήποτε είδους δεδομένων για την εξόρυξη δεδομένων. Εντούτοις, επιδεινώνονται για την ανάλυση καλαθιού νοικοκυράς επειδή αυτός ο τύπος ανάλυσης εξαρτάται σε μεγάλο βαθμό από συναλλαγές που γίνονται στα σημεία των πωλήσεων των προϊόντων.

Η ανάλυση καλαθιού νοικοκυράς έχει αποδειχθεί ιδιαίτερα χρήσιμη για δεδομένα μαζικών πωλήσεων όπως στα supermarkets, στα ψιλικατζίδικα, στα

φαρμακεία, και στα fast food, όπου πολλές από τις αγορές γίνονται παραδοσιακά με τα μετρητά. Οι συναλλαγές μετρητών είναι ανώνυμες, κάτι το οποίο σημαίνει ότι το κατάστημα δεν έχει καμία γνώση για τους συγκεκριμένους πελάτες επειδή δεν υπάρχει καμία πληροφορία που να προσδιορίζει τον πελάτη στη συναλλαγή. Για τις ανώνυμες συναλλαγές, οι μόνες πληροφορίες που είναι γνωστές για την αγορά είναι η ημερομηνία και ο χρόνος, η θέση του καταστήματος, ο ταμίας, τα αντικείμενα που αγοράζονται, οποιαδήποτε δελτία που εξαγοράζονται, και το ποσό των χρημάτων. Με την ανάλυση καλαθιού νοικοκυράς, ακόμη και αυτά τα περιορισμένα δεδομένα παράγουν ενδιαφέροντα και αξιοποιήσιμα αποτελέσματα.

Η αυξανόμενη χρήση των πιστωτικών καρτών, των χρεωστικών καρτών, και των καρτών των καταστημάτων (που προσφέρουν ιδιαίτερη μεταχείριση και προνομία στους κατόχους τους) έχει συνέπεια λιγότερες ανώνυμες συναλλαγές, που παρέχουν στους αναλυτές περισσότερες δυνατότητες για την απόκτηση πληροφοριών για τους πελάτες και τη συμπεριφορά τους κατά τη διάρκεια του χρόνου. Οι δημογραφικές πληροφορίες για τα άτομα και τις οικογένειες και οι τάσεις της αγοράς είναι διαθέσιμες στους αναλυτές για να βελτιώσουν περαιτέρω τα προφίλ των πελατών. Αυτές οι πρόσθετες πληροφορίες μπορούν να ενσωματωθούν στην ανάλυση χρησιμοποιώντας τα εικονικά στοιχεία (που περιγράφηκαν πιο πριν).

Σε άλλους τομείς όπως στις τραπεζικές εργασίες και την ιατρική φροντίδα, οι ανώνυμες συναλλαγές δεν ισχύουν. Όλες οι «συναλλαγές» με έναν πελάτη περιλαμβάνουν τον αριθμό λογαριασμού, την ταυτότητα ασθενούς, ή τα συναφή. Αυτό επιτρέπει τα ίδια δεδομένα να χρησιμοποιηθούν για την ανάλυση δεδομένων-χρόνου, την οποία θα περιγράψουμε όχι με πολλές λεπτομέρειες αργότερα σε αυτό το κεφάλαιο.

2.5 Παράγοντας τους κανόνες από όλα αυτά τα δεδομένα.

Υπολογίζοντας τον αριθμό των φορών που ένας δεδομένος συνδυασμός αντικειμένων εμφανίζεται στα δεδομένα συναλλαγής, είναι αποδεκτό , αλλά ένας συνδυασμός αντικειμένων δεν είναι ένας κανόνας. Μερικές φορές, ακριβώς αυτός ο συνδυασμός έχει ενδιαφέρον, όπως στο παράδειγμα με την πάνα, την μπύρα, και τις Πέμπτες. Αλλά σε άλλες περιπτώσεις, έχει περισσότερο νόημα να βρεθεί ένας ελλοχεύοντας κανόνας.

Τι είναι ένας κανόνας; Ένας κανόνας έχει δύο μέρη, έναν όρο και ένα αποτέλεσμα, και αντιπροσωπεύεται συνήθως ως δήλωση:

Εάν ο όρος ισχύει, οδηγεί σε ένα αποτέλεσμα.

Εάν ο κανόνας λέει, "Εάν υπηρεσία συνδιάσκεψης, τότε αναμονή κλήσης" τον ερμηνεύουμε ως: " Εάν αγοράζει κάποιος την υπηρεσία συνδιάσκεψης, αγοράζει και την αναμονή κλήσης."

Στην πράξη, οι πιο αξιοποιήσιμοι κανόνες έχουν μόνο ένα στοιχείο ως αποτέλεσμα. Έτσι, ένας κανόνας όπως

- Εάν πάνες και Πέμπτη, τότε μπύρα
- είναι πιο χρήσιμος από
- Εάν Πέμπτη, τότε πάνες και μπύρα.

Τα κατασκευάσματα όπως ο πίνακας των περιστατικών που συμβαίνουν μαζί παρέχουν πληροφορίες για το ποιοί συνδυασμοί αντικειμένων εμφανίζονται συνηθέστερα στις συναλλαγές. Χάριν της απεικόνισης, ας πούμε ότι ο πιο κοινός συνδυασμός έχει τρία στοιχεία, α, β, και γ. Οι μόνοι κανόνες για να εξετάσουν είναι εκείνοι και με τα τρία στοιχεία στον κανόνα και με ακριβώς ένα στοιχείο στο αποτέλεσμα:

- Αν α και β, τότε γ
- Αν α και γ, τότε β
- Αν β και γ, τότε α

Ο πίνακας 15 παρέχει ένα παράδειγμα, που παρουσιάζει τις πιθανότητες των αντικειμένων και των διάφορων συνδυασμών.

Πίνακας 15. Πιθανότητες των τριών αντικειμένων και οι συνδυασμοί τους

Συνδυασμός	Πιθανότητα
α	45%
β	42,50%
γ	40%
α και β	25%
β και γ	20%
α και γ	15%
α και β και γ	5%

Επειδή αυτοί οι τρεις κανόνες περιέχουν τα ίδια αντικείμενα, έχουν την ίδια υποστήριξη στα στοιχεία, 5 τοις εκατό. Τι γίνεται με το επίπεδο εμπιστοσύνης (confidence) τους; *Η εμπιστοσύνη είναι η αναλογία του αριθμού συναλλαγών με όλα τα αντικείμενα στον κανόνα προς τον αριθμό συναλλαγών με ακριβώς τα αντικείμενα που υπάρχουν στον όρο.* Η εμπιστοσύνη για τους τρεις κανόνες παρουσιάζεται στον πίνακα 16.

Πίνακας 16. Η εμπιστοσύνη στους κανόνες

Κανόνας	p(όρο)	p(όρος και αποτέλεσμα)	εμπιστοσύνη
Εάν α και β τότε γ	25%	5%	0,2
Εάν α και γ τότε β	20%	5%	0,25
Εάν β και γ τότε α	15%	5%	0,33

Τι λέει η εμπιστοσύνη πραγματικά; Λέγοντας ότι ο κανόνας "αν β και γ έπειτα α" έχει μια εμπιστοσύνη 0,33 είναι ισοδύναμος με το ότι όταν εμφανίζονται το β και το γ σε μια συναλλαγή, υπάρχει μια πιθανότητα 33 τοις εκατό ότι και το α εμφανίζεται επίσης σε αυτή. Δηλαδή μια φορά στις 3 το α εμφανίζεται με το β και το γ, και τις άλλες 2, το α όχι.

Ο βεβαιότερος κανόνας είναι και ο καλύτερος κανόνας, έτσι μπαίνουμε στον

πειρασμό να επιλέξουμε "αν β και γ τότε α" Αλλά υπάρχει ένα πρόβλημα. Αυτός ο κανόνας είναι πραγματικά χειρότερος από το να λέγαμε ακριβώς τυχαία ότι το α εμφανίζεται στη συναλλαγή. Το α εμφανίζεται σε 45 τοις εκατό των συναλλαγών αλλά ο κανόνας έχει μόνο 33 τοις εκατό εμπιστοσύνη. Ο κανόνας κάνει χειρότερα από το να υποθέτει τυχαία.

Αυτό προτείνει ένα άλλο μέτρο αποκαλούμενο *βελτίωση (improvement).* Η βελτίωση λέει πόσο καλύτερος ένας κανόνας είναι στην πρόβλεψη του αποτελέσματος από υποθέτοντας ακριβώς το αποτέλεσμα αρχικά. Δίνεται από τον ακόλουθο τύπο:

improvement = p(condition και result) / (PP (όρος) π (αποτέλεσμα))

Όταν η βελτίωση είναι μεγαλύτερη από 1, τότε ο προκύπτων κανόνας είναι καλύτερος στην πρόβλεψη του αποτελέσματος από την τυχαία πιθανότητα. Όταν είναι λιγότερο από 1, είναι χειρότερο. Ο ακόλουθος πίνακας 17. παρουσιάζει τη βελτίωση για τους τρεις κανόνες και για τον κανόνα με την καλύτερη βελτίωση.

Πίνακας 17. Μέτρηση βελτίωσης για 4 κανόνες

Κανόνας	Υποστήριξη	Εμπιστοσύνη	Βελτίωση
Εάν α κσι β τότε γ	5	0,2	0,5
Εάν α και γ τότε β	5	0,25	0,29
Εάν β και γ τότε α	5	0,33	0,74
Εάν α τότε β	25	0,59	1,31

Κανένας από τους κανόνες με 3 αντικείμενα δεν δείχνει καμία βελτίωση. Ο καλύτερος κανόνας στα δεδομένα στη πραγματικότητα έχει μόνο 2 αντικείμενα. Ο κανόνας «αν α τότε β» είναι 1.31 φορές καλύτερος στη πρόβλεψη από όταν το β είναι σε μια συναλλαγή που τυχαία μαντεύει. Σε αυτή τη περίπτωση, όπως και σε άλλες, ο καλύτερος κανόνας περιέχει λιγότερα αντικείμενα από ότι άλλοι κανόνες.

Όταν η βελτίωση είναι μεγαλύτερη από 1, το αποτέλεσμα παράγει έναν καλύτερο κανόνα. Εάν ο κανόνας

Εάν β και γ τότε α

έχει μια εμπιστοσύνη 0,33, τότε ο κανόνας

Εάν β και γ τότε ΟΧΙ α

έχει μια εμπιστοσύνη 0,67.

Δεδομένου ότι το Α εμφανίζεται σε 45 τοις εκατό των συναλλαγών, δεν εμφανίζεται σε 55 τοις εκατό τους. Η εφαρμογή του ίδιου μέτρου βελτίωσης δείχνει ότι η βελτίωση αυτού του νέου κανόνα είναι 1,22 (0.67/0.55). Ο αρνητικός κανόνας είναι χρήσιμος. Ο κανόνας "αν α και β τότε ΟΧΙ γ" έχει μια βελτίωση 1,33, καλύτερα από οποιονδήποτε από τους άλλους κανόνες.

Οι κανόνες παράγονται από τις βασικές πιθανότητες διαθέσιμες στον πίνακα περιστατικού που συμβαίνουν μαζί. Οι χρήσιμοι κανόνες έχουν μια βελτίωση που είναι μεγαλύτερη από 1. Όταν τα αποτελέσματα βελτίωσης είναι χαμηλά, μπορούν να αυξηθούν με την άρνηση των κανόνων. Εντούτοις, μπορούμε να διαπιστώσουμε ότι οι αρνούμενοι κανόνες δεν είναι τόσο χρήσιμοι όσο οι αρχικοί κανόνες συσχέτισης όταν φτάσουμε στο σημείο να ενεργήσουμε σύμφωνα με τα αποτελέσματα.

Επιπρόσθετα αναφέρονται εν' συντομία και οι κανόνες διαχωρισμού. Ένας *κανόνας διαχωρισμού* είναι παρόμοιος με έναν κανόνα συσχέτισης εκτός από το ότι μπορεί να έχει το συνδετήρα "ΚΑΙ ΟΧΙ" στον όρο εκτός από "ΚΑΙ". Ένας χαρακτηριστικός κανόνας διαχωρισμού μοιάζει με: *Εάν α και όχι β τότε γ*. Οι κανόνες διαχωρισμού μπορούν να παραχθούν από μια απλή προσαρμογή του βασικού αλγορίθμου της Ανάλυση Καλαθιού της Νοικοκυράς. Η προσαρμογή πρόκειται να εισαγάγει ένα νέο σύνολο στοιχείων που είναι τα αντίστροφα κάθε ένα από τα αρχικά στοιχεία. Κατόπιν, τροποποιείται κάθε συναλλαγή έτσι ώστε να περιλαμβάνει ένα αντίστροφο στοιχείο εάν, και μόνο εάν, δεν περιέχει το αρχικό στοιχείο.

2.6 Ξεπερνώντας τα πρακτικά όρια

Η παραγωγή των κανόνων ένωσης είναι μια πολλαπλών βημάτων διαδικασία. Ο γενικός αλγόριθμος είναι:

1.Παραγωγη της μήτρας περιστατικών που συμβαίνουν μαζί για τα μοναδικά αντικείμενα.

2. Παραγωγή της μήτρας περιστατικών που συμβαίνουν μαζί για δύο αντικείμενα. Χρησιμοποιούμε αυτό για να βρούμε τους κανόνες με δύο αντικείμενα.

3. Παραγωγή της μήτρας περιστατικών που συμβαίνουν μαζί για τρία αντικείμενα. Χρησιμοποιούμε αυτό για να βρούμε τους κανόνες με τρία αντικείμενα.

4.Ομοιως τα παραπάνω βήματα και για περισσότερα αντικείμενα.

Παραδείγματος χάριν, το παντοπωλείο που πουλάει χυμό από πορτοκάλι, γάλα, απορρυπαντικό, σόδα, και καθαριστικό τζαμιών, το πρώτο βήμα υπολογίζει τις αριθμήσεις για κάθε ένα από αυτά τα αντικείμενα. Κατά τη διάρκεια του δεύτερου βήματος, οι ακόλουθες αριθμήσεις δημιουργούνται:

- χυμός από πορτοκάλι και γάλα, χυμός από πορτοκάλι και απορρυπαντικό, χυμός από πορτοκάλι και σόδα, χυμός από πορτοκάλι και καθαριστικό.
- γάλα και απορρυπαντικό, γάλα και σόδα, γάλα και καθαριστικό
- απορρυπαντικό και σόδα, απορρυπαντικό και καθαριστικό
- Σόδα και καθαριστικό

Αυτά είναι συνολικά 10 αριθμήσεις. Το τρίτο πέρασμα παίρνει όλους τους συνδυασμούς τριών αντικειμένων και τα λοιπά. Φυσικά, κάθε ένα από αυτά τα στάδια μπορεί να απαιτήσει ένα χωριστό πέρασμα μέσω των αντικειμένων ή τα πολλαπλάσια στάδια μπορούν να συνδυαστούν σε ένα ενιαίο πέρασμα με το να εξετάσουν τους διαφορετικούς αριθμούς συνδυασμών συγχρόνως.

Αν και δεν είναι προφανές όταν υπάρχουν ακριβώς πέντε στοιχεία, η αύξηση του αριθμού αντικειμένων στους συνδυασμούς απαιτεί εκθετικά περισσότερους υπολογισμούς. Αυτό οδηγεί σε εκθετική αύξηση των εκτέλεσης και μακρές αναμονές κατά την εξέταση των συνδυασμών με περισσότερα από τρία ή τέσσερα στοιχεία. Η λύση είναι η *περικοπή (prunning)*. Η περικοπή είναι μια τεχνική μείωσης του αριθμού των αντικειμένων και των συνδυασμών των αντικειμένων σε κάθε βήμα. Σε κάθε στάδιο, ο αλγόριθμος ρίχνει έξω ορισμένους συνδυασμούς που δεν ικανοποιούν κάποια κριτήρια κατώτατων ορίων.

Ο πιο κοινός μηχανισμός περικοπής καλείται *ελάχιστη περικοπή υποστήριξης (minimum support pruning)*. Υπενθυμίζουμε ότι η υποστήριξη αναφέρεται στον αριθμό συναλλαγών στη βάση δεδομένων όπου ο κανόνας ισχύει. Η ελάχιστη περικοπή υποστήριξης απαιτεί ότι ένας κανόνας ισχύει σε έναν ελάχιστο αριθμό συναλλαγών. Παραδείγματος χάριν, εάν υπάρχουν 1 εκατομμύριο συναλλαγές και η ελάχιστη υποστήριξη είναι 1 τοις εκατό, τότε μόνο οι κανόνες που υποστηρίζονται από 10.000 συναλλαγές είναι ενδιαφέροντες. Αυτό έχει νόημα, επειδή ο σκοπός που παράγονται οι κανόνες είναι να ακολουθηθεί κάποιο είδος δράσης, όπως τοποθέτηση των πανών στον ίδιο διάδρομο με τη μπύρα, και η δράση πρέπει να έχει επιπτώσεις σε αρκετές συναλλαγές για να είναι σημαντική.

Ο ελάχιστος περιορισμός υποστήριξης έχει μια ιδιαίτερη επίδραση. Ας υποθέσουμε ότι εξετάζουμε έναν κανόνα με τέσσερα στοιχεία σε αυτόν, όπως

Εάν α, β, και γ, τότε δ.

Χρησιμοποιώντας την ελάχιστη περικοπή υποστήριξης, αυτός ο κανόνας πρέπει να ισχύει σε τουλάχιστον 10.000 συναλλαγές στα δεδομένα.

Ακολουθεί ότι:

- Το α πρέπει να εμφανιστεί σε τουλάχιστον 10.000 συναλλαγές και,
- το β πρέπει να εμφανιστεί σε τουλάχιστον 10.000 συναλλαγές και,
- το γ πρέπει να εμφανιστεί σε τουλάχιστον 10.000 συναλλαγές και,
- το δ πρέπει να εμφανιστεί σε τουλάχιστον 10.000 συναλλαγές.

Με άλλα λόγια, η ελάχιστη περικοπή υποστήριξης αποβάλλει τα στοιχεία που δεν εμφανίζονται σε αρκετές συναλλαγές! Υπάρχουν δύο τρόποι να γίνει αυτό. Ο πρώτος τρόπος είναι να αποβληθούν τα στοιχεία από την επεξεργασία. Ο δεύτερος τρόπος είναι να χρησιμοποιηθεί η ταξονομία για να γενικεύσει τα στοιχεία, έτσι τα προκύπτοντα γενικευμένα στοιχεία ικανοποιούν το κριτήριο κατώτατων ορίων. Το κριτήριο κατώτατων ορίων ισχύει για κάθε βήμα στον αλγόριθμο. Το ελάχιστο κατώτατο όριο επίσης υπονοεί ότι:

- Το α και το β πρέπει να εμφανιστούν μαζί σε τουλάχιστον 10.000 συναλλαγές και,
- α και γ πρέπει να εμφανιστούν μαζί σε τουλάχιστον 10.000 συναλλαγές και,
- α και δ πρέπει να εμφανιστούν μαζί σε τουλάχιστον 10.000 συναλλαγές,
- Και ούτω καθ΄ εξής.

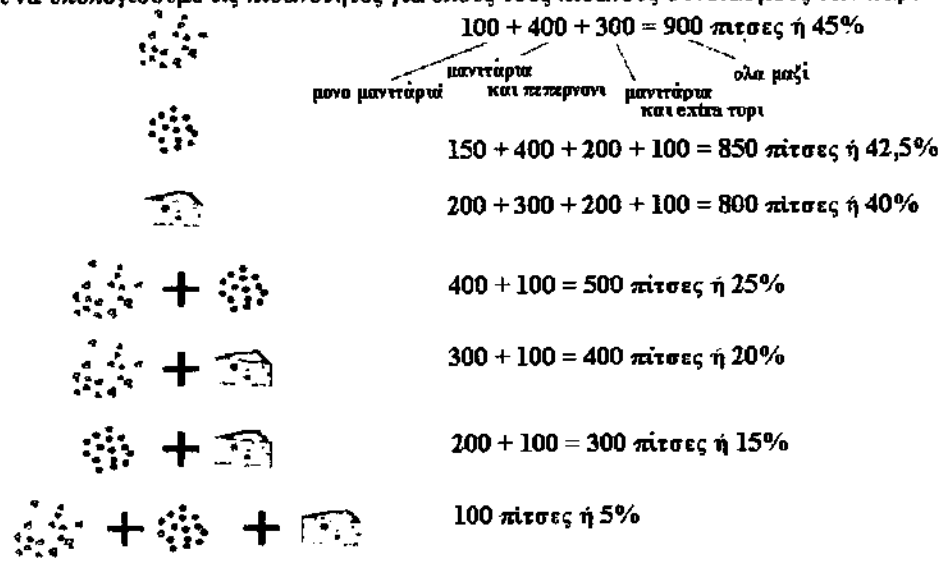
Κάθε βήμα του υπολογισμού του πίνακα περιστατικών που συμβαίνουν μαζί μπορεί να αποβάλει τους συνδυασμούς αντικειμένων που δεν ικανοποιούν το κατώτατο όριο, μειώνοντας το μέγεθός του και τον αριθμό συνδυασμών που επεξεργάζονται κατά τη διάρκεια του επόμενου περάσματος.

Το σχήμα 18 είναι ένα παράδειγμα για το πώς ο υπολογισμός πραγματοποιείται. Σε αυτό το παράδειγμα, επιλέγοντας ένα κατώτατο επίπεδο υποστήριξης 10 τοις εκατό δεν θα λάμβανε υπ' όψιν όλους τους συνδυασμούς με τρία αντικείμενα, καθώς και τους σχεσιακούς κανόνες τους. Αυτό είναι ένα παράδειγμα όπου η περικοπή δεν έχει επίδραση στον καλύτερο κανόνα δεδομένου ότι ο καλύτερος κανόνας έχει μόνο δύο αντικείμενα. Στην περίπτωση της πίτσας, αυτά τα επιπρόσθετα είναι όλα αρκετά κοινά, έτσι δεν περικόπτονται χωριστά. Εάν οι αντσούγιες περιλήφθηκαν στην ανάλυση, και υπάρχουν μόνο 15 από τις 2,000 πίτσες που περιέχουν αντσούγιες, τότε μια ελάχιστη υποστήριξη 10 τοις εκατό, ή ακόμα και 1 τοις εκατό, θα απέβαλλε τις αντσούγιες κατά τη διάρκεια του πρώτου περάσματος.

Η καλύτερη επιλογή για την ελάχιστη υποστήριξη εξαρτάται από τα στοιχεία και την κατάσταση. Είναι επίσης δυνατό να ποικίλει η ελάχιστη υποστήριξη καθώς ο αλγόριθμος προχωρεί. Παραδείγματος χάριν, χρησιμοποιώντας διαφορετικά επίπεδα σε διαφορετικά στάδια μπορούν να βρεθούν ασυνήθιστοι συνδυασμοί κοινών αντικειμένων (μειώνοντας το επίπεδο υποστήριξης για διαδοχικά βήματα) ή σχετικά κοινοί συνδυασμοί ασυνήθιστων στοιχείων (με την αύξηση του επιπέδου υποστήριξης). Η ποικιλία της ελάχιστης υποστήριξης βοηθά να βρεθούν οι αξιοποιήσιμοι κανόνες, έτσι οι κανόνες που παράγονται δεν είναι όλοι όπως η διαπίστωση ότι το φυστικό-βούτυρο και το ζελέ αγοράζονται συχνά από το κοινό.

Μια πιτσαρία εχει πουλήσει 2,000 πίτσες απο τις οποίες:
100 με μανιτάρια, 150 πεπερόνι, 200 με έξτρα τυρί
400 με μανιτάρια και πεπερόνι, 300 με μανιτάρια και έξτρα τυρί, 200 με πεπερόνι και έξτρα τυρί
100 με μανιτάρια, πεπερόνι και έξτρα τυρί
550 δεν ειχαν επιπρόσθετα

Πρέπει να υπολογίσουμε τις πιθανότητες για όλους τους πιθανούς συνδιασμούς των παραπάνω αντικειμένων



Σχήμα 18: Μετρώντας τις συχνότητες των πωλήσεων πίτσας

2.7 Το πρόβλημα των μεγάλων δεδομένων

Ένα χαρακτηριστικό εστιατόριο fast food προσφέρει αρκετά αντικείμενα στο μενού του, για παράδειγμα ας υποθέσουμε ότι υπάρχουν 100. Για να χρησιμοποιηθούν οι πιθανότητες για να παραγάγουν τους κανόνες συσχέτισης, οι αριθμήσεις πρέπει να υπολογιστούν για κάθε συνδυασμό αντικειμένων. Ο αριθμός συνδυασμών ενός δεδομένου μεγέθους τείνει να αυξάνεται εκθετικά. Ένας συνδυασμός με τρία στοιχεία να είναι τηγανητές πατάτες μικρού μεγέθους, cheeseburger, και κόκα κόλα διαίτης μεσαίου μεγέθους. Σε ένα μενού με 100 αντικείμενα, πόσοι συνδυασμοί υπάρχουν με τρία αντικείμενα του μενού; Υπάρχουν 161.700! Ο πίνακας 19 επιδεικνύει πόσο γρήγορα ο αριθμός συνδυασμών αυξάνεται.

Πίνακας 19. Ο αριθμός των συνδυασμών των αντικειμένων αυξάνεται γρήγορα

Αριθμός αντικειμένων στον συνδυασμό	Αριθμός συνδυασμών
1	100
2	4,950
3	161,700
4	3,921,225
5	75,287,520
6	1,192,052,400
7	16,007,560,800
8	186,087,894,300

Εξάλλου, ένα χαρακτηριστικό supermarket έχει τουλάχιστον 10.000 διαφορετικά αντικείμενα στοκ, η ακόμα και 20.000 ή 30.000. Υπολογίζοντας την υποστήριξη, την εμπιστοσύνη, και την βελτίωση, οι αριθμοί ξεφεύγουν γρήγορα από τον έλεγχο καθώς ο αριθμός αντικειμένων που εμπλέκονται σε συνδυασμούς αυξάνεται. Υπάρχουν σχεδόν 50 εκατομμύρια πιθανοί συνδυασμοί δύο αντικειμένων στο Supermarket και πάνω από 100 δισεκατομμύρια συνδυασμοί τριών αντικειμένων. Αν και οι υπολογιστές γίνονται όλο και γρηγορότεροι και φτηνότεροι, είναι ακόμα πολύ δύσκολο να υπολογιστούν οι αριθμήσεις για αυτόν τον αριθμό συνδυασμών. Ο υπολογισμός των αριθμήσεων για πέντε ή περισσότερα στοιχεία είναι απαγορευτικά «ακριβός». Η χρήση των ταξονομιών μειώνει τον αριθμό αντικειμένων σε ένα εύχρηστο μέγεθος.

Ο αριθμός συναλλαγών είναι επίσης πολύ μεγάλος. Κατά τη διάρκεια ενός έτους, μια αλυσίδα καταστημάτων παγκόσμιας εμβέλειας θα παραγάγει δεκάδες εκατομμυρίων συναλλαγών. Κάθε μια από αυτές τις συναλλαγές αποτελείται από ένα ή περισσότερα αντικείμενα. Έτσι, ο καθορισμός εάν ένας ιδιαίτερος συνδυασμός αντικειμένων είναι παρών σε μια συγκεκριμένη συναλλαγή μπορεί να απαιτήσει αρκετή προσπάθεια, φανταστείτε να γίνει αυτό για όλες τις συναλλαγές.

2.8 Ανάλυση χρονοσειρών

Η Ανάλυση Καλαθιού της Νοικοκυράς αναλύει τα γεγονότα που συμβαίνουν την ίδια στιγμή, ποια αντικείμενα αγοράζονται σε μία δεδομένη στιγμή. Η επόμενη λογική ερώτηση αφορά τις ακολουθίες γεγονότων και τι σημαίνουν. Τα παραδείγματα των αποτελεσμάτων σε αυτήν την περιοχή είναι:

- Νέοι οι ιδιοκτήτες σπιτιού αγοράζουν τις κουρτίνες στο ντους πριν αγοράσουν τα έπιπλα.
- Όταν ένας πελάτης πηγαίνει σε ένα υποκατάστημα τράπεζας τακτοποιεί τον λογαριασμού του, υπάρχει μια καλή πιθανότητα ότι θα κλείσει όλους τους λογαριασμούς του.

Τα δεδομένα των χρονοσειρών απαιτούν συνήθως κάποιο τρόπο αναγνώρισης του πελάτη. Οι ανώνυμες συναλλαγές δεν μπορούν να αποκαλύψουν ότι οι νέοι ιδιοκτήτες σπιτιών αγοράζουν τις κουρτίνες του ντους προτού να αγοράσουν τα έπιπλα. Αυτό απαιτεί την «παρακολούθηση» του κάθε πελάτη, καθώς επίσης και ποιοι πελάτες αγόρασαν πρόσφατα ένα σπίτι. Δεδομένου ότι οι μεγαλύτερες αγορές γίνονται συχνά με τις πιστωτικές κάρτες ή τις χρεωστικές κάρτες, αυτό δεν δημιουργεί κανένα πρόβλημα. Για τα προβλήματα σε άλλους τομείς, όπως η έρευνα των επιπτώσεων των ιατρικών περιθάλψεων ή της συμπεριφοράς πελατών μέσα σε μια τράπεζα, όλες οι συναλλαγές περιλαμβάνουν πληροφορίες ταυτοποίησης και αναγνώρισης του πελάτη.

Προκειμένου να εξεταστούν οι αναλύσεις των χρονοσειρων στους πελάτες, πρέπει να υπάρξει κάποιος τρόπος αναγνώρισης των πελατών κατά τη διάρκεια του χρόνου. Χωρίς έναν τρόπο παρακολούθησης συγκεκριμένων πελατών, δεν υπάρχει κανένας λόγος να αναλυθεί η συμπεριφορά τους.

Δεχόμαστε ότι μια χρονοσειρά είναι μια διαταγμένη ακολουθία αντικειμένων. Διαφέρει από μια συναλλαγή μόνο στη διάταξη. Γενικά, η χρονοσειρά περιέχει τον προσδιορισμό των πληροφοριών για τον πελάτη, δεδομένου ότι αυτές οι πληροφορίες χρησιμοποιούνται για να συνδέσουν τις διαφορετικές συναλλαγές σε μια σειρά. Υπάρχουν πολλές τεχνικές για τη ανάλυση μιας χρονοσειρά, όπως τα νευρωνικά

δίκτυα, αλλά ξεφεύγει μάλλον από τους σκοπούς αυτού του κεφαλαίου όποτε θα παραληφθεί.

2.9 Πλεονεκτήματα των Κανόνων Συσχέτισης

Τα πλεονεκτήματα της μεθόδου είναι:

• Παράγει σαφή και κατανοητά αποτελέσματα.

Τα αποτελέσματα που παράγει η Ανάλυση Καλαθιού της Νοικοκυράς είναι κανόνες συσχέτισης (association rules). Αυτοί εκφράζονται εύκολα στα αγγλικά ή ως προτάσεις SQL. Η έκφραση των πρότυπων στα αντικείμενα ως "εάν - τότε" κανόνες, καθιστούν τα αποτελέσματα κατανοητά και διευκολύνουν τη μετατροπή των αποτελεσμάτων σε ενέργειες. Σε μερικές περιπτώσεις, μόνο το σύνολο σχετικών αντικειμένων είναι ενδιαφέρον και οι κανόνες δεν πρέπει ακόμη και να παραχθούν.

• Υποστηρίζει την μη κατευθυνόμενη (undirected) εξόρυξη δεδομένων.

Η μη κατευθυνόμενη εξόρυξη δεδομένων είναι πολύ σημαντική κατά την προσέγγιση ενός μεγάλου συνόλου δεδομένων. Η ανάλυση καλαθιού αγοράς είναι μια κατάλληλη τεχνική, όταν μπορεί να εφαρμοστεί, για να αναλύσει τα στοιχεία και για να γίνει έτσι μια αρχή για περαιτέρω επεξεργασία. Οι περισσότερες τεχνικές εξόρυξης δεδομένων δεν χρησιμοποιούνται πρώτιστα για την undirected εξόρυξη δεδομένων. Η Ανάλυση Καλαθιού της Νοικοκυράς, αφ' ετέρου, χρησιμοποιείται σε αυτήν την περίπτωση και παρέχει τα σαφή αποτελέσματα.

• Οι υπολογισμοί που χρησιμοποιεί είναι απλοί στη κατανόηση.

Οι υπολογισμοί που απαιτούνται για να εφαρμόσουν την Ανάλυση Καλαθιού της Νοικοκυράς είναι μάλλον απλοί, αν και ο αριθμός υπολογισμών αυξάνεται πολύ γρήγορα ανάλογα με τον αριθμό συναλλαγών και τον αριθμό διαφορετικών αντικειμένων στην ανάλυση. Τα μικρότερα προβλήματα μπορούν να επεξεργαστούν κάνοντας διάφορους υπολογισμούς σε ένα λογιστικό φύλλο (spreadsheet). Αυτό καθιστά την τεχνική αυτή πιο άνετη στη

χρήση από άλλες σύνθετες τεχνικές, όπως τους γενετικούς αλγορίθμους ή τα νευρωνικά δίκτυα.

2.10 Αδυναμίες της μεθόδου

Οι αδυναμίες της μεθόδου της ανάλυσης καλαθιού νοικοκυράς είναι:

- απαιτεί περισσότερη υπολογιστική δύναμη (που αυξάνεται με εκθετικούς ρυθμούς) καθώς το μέγεθος των δεδομένων του προβλήματος αυξάνεται.
- έχει μια περιορισμένη υποστήριξη για τις ιδιότητες των αντικειμένων.
- είναι δύσκολο να καθοριστεί ο σωστός αριθμός των στοιχείων.
- σπάνια στοιχεία δεν λαμβάνονται υπ' οψιν.

Αναλυτικότερα:

- **Εκθετική αύξηση απαιτήσεων όταν αυξάνεται το μεγέθους του προβλήματος**

Οι υπολογισμοί που απαιτούνται για να παραγάγουν τους κανόνες συσχέτισης αυξάνονται με εκθετικούς ρυθμούς ανάλογα φυσικά με τον αριθμό των αντικειμένων και την πολυπλοκότητα της εξέτασης των κανόνων. Η λύση είναι να μειωθεί ο αριθμός αντικειμένων με γενικεύοντας τους. Εντούτοις, τα γενικότερα αντικείμενα είναι συνήθως λιγότερο αξιοποιήσιμα. Οι μέθοδοι για να ελέγξουν τον αριθμό υπολογισμών, όπως η ελάχιστη περικοπή υποστήριξης (minimum support pruning), μπορούν να παραλείψουν σημαντικούς κανόνες από το να ληφθούν υπ' όψιν.

- **Περιορισμένη υποστήριξη για τις ιδιότητες των δεδομένων**

Η Ανάλυση Καλαθιού της Νοικοκυράς είναι μια τεχνική που ειδικεύεται στα αντικείμενα μιας συναλλαγής. Τα αντικείμενα υποτίθεται ότι είναι ίδια εκτός από ένα χαρακτηριστικό που τα προσδιορίζει, όπως ο τύπος προϊόντων. Όπου μπορεί να εφαρμοστεί, η μέθοδος αυτή είναι πολύ ισχυρή. Εντούτοις, δεν έγκειται όλα τα προβλήματα αυτήν την περιγραφή. Η χρήση των ταξινομιών στοιχείων και των εικονικών αντικειμένων καθιστούν τους κανόνες πιο εκφραστικούς και ευκολότερους στη χρήση.

• **Καθορισμός των σωστών στοιχείων**

Πιθανώς το δυσκολότερο πρόβλημα κατά τον εφαρμογή της μεθόδου είναι ο καθορισμός του σωστό συνόλου αντικειμένων που χρησιμοποιούνται στην ανάλυση. Με το να γενικεύσει τα αντικείμενα η ταξονομία τους, μπορεί να εξασφαλισθεί ότι οι συχνότητες των αντικειμένων που χρησιμοποιούνται στην ανάλυση είναι σχεδόν ίδιες. Αν και αυτή η διαδικασία γενίκευσης χάνει κάποιες πληροφορίες, τα εικονικά στοιχεία μπορούν έπειτα να επανεισαχθούν στην ανάλυση για να βρεθούν έτσι οι πληροφορίες που εκτείνονται στα γενικευμένα αντικείμενα.

• **Η μέθοδος έχει πρόβλημα με αντικείμενα που εμφανίζονται σπάνια.**

Οι εργασίες της Ανάλυσης Καλαθιού της Νοικοκυράς λειτουργούν καλύτερα όταν έχουν περίπου όλα τα αντικείμενα την ίδια συχνότητα στα δεδομένα. Τα αντικείμενα που εμφανίζονται σπάνια βρίσκονται σε πολύ λίγες συναλλαγές και θα απορριφθούν. Η τροποποίηση του ελάχιστου κατώτατου ορίου υποστήριξης για να ληφθεί υπ' όψιν η αξία προϊόντων είναι ένας τρόπος να εξασφαλιστεί ότι τα ακριβά αντικείμενα παραμένουν, ακόμα κι αν μπορούν να εμφανίζονται σπάνια στα δεδομένα. Η χρήση των ταξονομιών στοιχείων μπορεί να εξασφαλίσει ότι τα σπάνια αντικείμενα περιλαμβάνονται στην ανάλυση με κάποια μορφή.

2.11 Πότε εφαρμόζεται η μέθοδος.

Η Ανάλυση Καλαθιού της Νοικοκυράς εφαρμόζεται σε μη κατευθυνόμενα (undirected) προβλήματα εξόρυξης δεδομένων που αποτελούνται από καθορισμένα με σαφήνεια αντικείμενα που ομαδοποιούνται με ενδιαφέροντες τρόπους. Αυτά τα προβλήματα εμφανίζονται συνήθως στον τομέα των λιανικών πωλήσεων όπου οι συναλλαγές προϊόντων είναι η βάση για την ανάλυση. Τα παρόμοια προβλήματα μπορούν να βρεθούν σε άλλους τομείς.

Η μέθοδος μπορεί επίσης να εφαρμοστεί σε μερικά κατευθυνόμενα προβλήματα εξόρυξης δεδομένων σε αυτούς τους τομείς. Μπορεί να τεθεί σε εφαρμογή πάνω σε ένα καθορισμένο με σαφήνεια υποσύνολο των συναλλαγών, όπως οι συναλλαγές από τα νέα καταστήματα ή τα φάρμακα που γράφονται από

παθολόγους σε ασθενείς από διάφορα ταμεία, έτσι ώστε να βρεθούν κάποια παραρτήματα σε ένα υποσύνολο ενδιαφέροντος. Ο βασικός αλγόριθμος μπορεί επίσης να τροποποιηθεί για να εξετάσει μόνο τους κανόνες που περιέχουν ένα ιδιαίτερο αντικείμενο, όπως ένα νέο προϊόν, έτσι ώστε να γίνουν περισσότερο κατανοητά τα πρότυπα των πωλήσεων.

Προβλήματα που έχουν σχέση με το χρόνο είναι μια άλλη περιοχή όπου αυτές οι μέθοδοι μπορούν να εφαρμοστούν. Πολλά τέτοια προβλήματα μπορούν να προσαρμοστούν για την μέθοδο με σχετικά απλούς μετασχηματισμούς των δεδομένων στη χρονική σειρά.

Στο επόμενο κεφάλαιο αναφέρονται οι πιο γνωστοί αλγόριθμοι παραγωγής κανόνων συσχέτισης

3. ΑΛΓΟΡΙΘΜΟΙ ΠΑΡΑΓΩΓΗΣ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ

Οι στόχοι της Market-Basket εξερεύνησης είναι:

1. ***Κανόνες συσχέτισης***. Εχουν τη μορφή {x1, x2,...,xn}->y. Αυτό σημαίνει ότι αν στο ίδιο καλάθι έχουμε τα x1, x2,...,xn, τότε υπάρχει μεγάλη πιθανότητα να βρούμε το y. Η πιθανότητα εύρεσης του y καλείται "*confidence*", του κανόνα. Ψάχνουμε, κυρίως, κανόνες που η πιθανότητά τους είναι πάνω από κάποιο κατώφλι. Δηλαδή, κανόνες με σημαντικά υψηλότερο "confidence", από ό,τι θα ήταν αν τα αντικείμενα τοποθετούνται τυχαία μέσα στο καλάθι. Ακόμη, μπορεί να βρίσκαμε έναν κανόνα {milk, butter}->bread, απλώς γιατί πολλοί άνθρωποι αγοράζουν ψωμί.
2. ***Causality***. Ιδανικά, θέλουμε να ξέρουμε ότι σε έναν κανόνα συσχέτισης η παρουσία των x1 και x2 συνεπάγει τη παρουσία του y. Π.χ. αν μειώσουμε τις τιμές των πάνων κι αυξήσουμε τις τιμές της μπύρας μπορούμε να δελεάσουμε τους πελάτες να αγοράζουν μπύρα ενώ αγοράζουν πάνες. Αυτό συμβαίνει γιατί "diapers cause beer".
3. ***Συχνά σύνολα αντικειμένων***. Σε πολλές περιπτώσεις ενδιαφερόμαστε για σύνολα αντικειμένων που εμφανίζονται συχνά στο καλάθι. Ξεκινάμε με την υπόθεση ότι μας ενδιαφέρουν σύνολα αντικειμένων με μεγάλο βαθμό υποστήριξης (*support*), δηλαδή αντικείμενα να εμφανίζονται μαζί σε πολλά καλάθια. Βρίσκουμε σύνολα {x1, x2,...,xn, y} που εμφανίζονται σε σημαντικό ποσοστό καλαθιών (*κατώφλι υποστήριξης*).

Χρησιμοποιούμε τον όρο "*συχνά σύνολα αντικειμένων*" για ένα σύνολο S που εμφανίζεται τουλάχιστο σε ένα ποσοστό s των καλαθιών, όπου s είναι μια σταθερά της τάξης του 1%. Τα δεδομένα είναι πάρα πολλά για την κυρία μνήμη. Είτε είναι αποθηκευμένα σε μια RDB ως συσχέτιση Baskets(BID, item) είτε ως αρχείο flat (BID, item1, item2,..., itemn) όταν εκτελούμε το χρόνο εκτέλεσης του αλγορίθμου:

- Μετράμε τον αριθμό προσπελάσεων στα δεδομένα. Αφού το κύριο κόστος είναι ο χρόνος διαβάσματος των δεδομένων από το δίσκο, το πόσες φορές θα

χρειαστεί να διαβάσουμε τα δεδομένα είναι μια καλή μέτρηση του χρόνου εκτέλεσης του αλγορίθμου.

- Υπάρχει μια έννοια-κλειδί, που καλείται monotonicity ή a-priori trick που μας βοηθά να βρούμε συχνά σύνολα αντικειμένων. Αν ένα σύνολο αντικειμένων S είναι συχνό (εμφανίζεται τουλάχιστο σε ένα ποσοστό s των καλαθιών), τότε κάθε υποσύνολο του S είναι επίσης συχνό.

Για να βρούμε συχνά σύνολα αντικειμένων:

1. Προχωράμε επίπεδο-επίπεδο (*levelwise*). Βρίσκουμε πρώτα τα συχνά αντικείμενα (σύνολα μεγέθους 1) μετά βρίσκουμε τα συχνά ζευγάρια, τις συχνές τριάδες κοκ. Εδώ θα εστιάσουμε στην εύρεση συχνών ζευγαριών, διότι:
    a. Συχνά τα ζευγάρια μας αρκούν
    b. Σε πολλά σύνολα δεδομένων, το δυσκολότερο κομμάτι είναι η εύρεση ζευγαριών. Προχωρώντας σε υψηλότερα επίπεδα (δηλ., τριάδες κλπ) παίρνει λιγότερο χρόνο από την εύρεση συχνών ζευγαριών.

    Οι levelwise αλγόριθμοι κάνουν ένα πέρασμα για κάθε επίπεδο.

2. Βρίσκουμε σε ένα ή σε λίγα περάσματα, τα πιο συχνά σύνολα αντικειμένων, έτσι ώστε κανένα υπερσύνολο του S να μην είναι συχνό.

Ο **A-Priori** αλγόριθμος (Agrawal και Srikant, 1994) προχωρά επίπεδο-επίπεδο (*levelwise*).

1. Δεδομένου ενός κατωφλίου υποστήριξης s, στο πρώτο πέρασμα βρίσκουμε τα αντικείμενα που εμφανίζονται τουλάχιστο σε ένα ποσοστό s των καλαθιών. Αυτό το σύνολο το ονομάζουμε L1. Ενδεχόμενα, υπάρχει αρκετή κύρια μνήμη να μετράμε τις εμφανίσεις κάθε αντικειμένου, αφού ένα τυπικό κατάστημα πουλά περισσότερα από 100,000 διαφορετικά αντικείμενα.
2. Τα ζεύγη των αντικειμένων του L1 είναι τα υποψήφια ζεύγη C2 για το δεύτερο πέρασμα. Τα ζεύγη του C2 που περνάν το κατώφλι s είναι τα συχνά ζεύγη του L2.

3. Οι υποψήφιες τριάδες C3 είναι τα σύνολα {A,B,C} έτσι ώστε όλα τα {A,B}{A,C}{B,C} να βρίσκονται στο L2. Στο τρίτο πέρασμα μετράμε την εμφάνιση των τριάδων του C3 κι όσες περνούν το s είναι οι συχνές τριάδες L3.

3.1 Ο αλγόριθμος A Priori

Σε μια εμπορική συναλλαγή για παράδειγμα μπορούμε να θεωρήσουμε ότι A=1 σημαίνει ότι το προϊόν περιλαμβάνεται στο καλάθι των αγορών και αντίστοιχα η τιμή A=0 ότι το A δεν περιλαμβάνεται στην αγορά. Η πιθανότητα p(A=1,B=1,C=1) ονομάζεται *στήριξη* (support) ή *διάδοση* (prevalence) του συνδέσμου και εκφράζει την συχνότητα με την οποία ένας συγκεκριμένος σύνδεσμος εμφανίζεται στη βάση δεδομένων. Αντίστοιχα, η πιθανότητα p = p(C =1 | A =1, B =1) είναι η δεσμευμένη πιθανότητα να είναι C=1 δεδομένου ότι είναι A=1 και B=1. Η δεσμευμένη πιθανότητα p ονομάζεται, στην ορολογία της όρυξης δεδομένων, *ακρίβεια* (accuracy) ή *εμπιστοσύνη* (confidence) του συνδέσμου και εκφράζει τη σχετική συχνότητα εμφάνισης των στοιχείων και των συνδυασμών τους και ορίζεται ως εξής:

$$\text{ακρίβεια}[(A=1, B=1) \Rightarrow C=1] = p(A=1, B=1, C=1) / p(A=1, B=1)$$

Τυπικά, ο στόχος της ανακάλυψης συνδέσμων είναι να εντοπισθούν όλοι οι σύνδεσμοι που έχουν ακρίβεια μεγαλύτερη από ένα κατώφλι p_a και στήριξη μεγαλύτερη από p_s. Οι εξαγόμενοι από μια τέτοια διαδικασία κανόνες συνιστούν μια σχετικά υποτυπώδη μορφή γνώσης, αφού στην πραγματικότητα δίδουν μόνο μια περιορισμένη εικόνα για τις συνεμφανίσεις κάποιων γεγονότων σε ένα σύνολο δεδομένων (π.χ. προϊόντα που αγοράζονται συχνά μαζί) και δεν αναδεικνύουν ισχυρές καθολικές ιδιότητες για το σύνολο των δεδομένων. Έτσι οι σύνδεσμοι είναι πραγματικοί κανόνες (δεν δίδουν δηλαδή μια αιτιολογική ερμηνεία) αλλά παρατηρούμενες σχέσεις σε ένα συγκεκριμένο σύνολο δεδομένων. Δεν υπάρχουν στέρεοι τρόποι ελέγχου αυτών των κανόνων ώστε να αποτελέσουν εργαλεία πρόβλεψης. Η ανάλυση συνδέσμων και όποια συμπερασματολογία βασίζεται στην ανάλυση αυτή, στηρίζεται στην παραδοχή ότι συμπεριφορές του παρελθόντος επαναλαμβάνονται και στο μέλλον.

Σύμφωνα με τη γενική τυπολογία χαρακτηριστικών των αλγορίθμων όρυξης (βλ. προηγούμενα), ο γενικός αλγόριθμος a priori έχει ως αποστολή (task) την περιγραφή συνδέσμων μεταξύ των μεταβλητών. Η δομή του (structure) είναι οι κατά

πιθανότητα κανόνες συνδέσμων, ως δε συνάρτηση μέτρου (score function) έχει τα κατώφλια ακρίβειας και στήριξης. Η αναζήτηση γίνεται συστηματικά (breadth-first with pruning) ενώ η προσπέλαση των δεδομένων γίνεται με πολλαπλές γραμμικές σαρώσεις όλου του πίνακα δεδομένων.

Η συνάρτηση μέτρου που χρησιμοποιείται στην αναζήτηση συνδέσμων είναι μια απλή δυαδική συνάρτηση (δηλαδή συνάρτηση που λαμβάνει μόνο τις τιμές 0 και 1). Ορίζονται δύο κατώφλια: το p_s, που είναι ένα κάτω φράγμα για το επίπεδο στήριξης του συνδέσμου και το p_a, που είναι ένα κάτω φράγμα για το επίπεδο ακρίβειας του συνδέσμου. Για παράδειγμα, αν ενδιαφερόμαστε για τους συνδέσμους που καλύπτουν τουλάχιστον το 10% των περιπτώσεων (εγγραφών) ορίζουμε p_s = 0.1. Αν επίσης ενδιαφερόμαστε για συνδέσμους που είναι ακριβείς σε ποσοστό τουλάχιστον 90% ορίζουμε p_a =0.9. Για έναν οποιοδήποτε σύνδεσμο, η συνάρτηση μέτρου λαμβάνει την τιμή 1 αν ικανοποιούνται και οι δύο συνθήκες για την στήριξη και την ακρίβεια που ορίζονται με τα παραπάνω κατώφλια. Διαφορετικά η συνάρτηση μέτρου λαμβάνει την τιμή 0. Στόχος λοιπόν αυτής της τεχνικής όρυξης είναι εύρεση όλων των συνδέσμων (προτύπων) με μέτρο 1.

Σε ότι αφορά την αποδοτικότητα της διαδικασίας αναζήτησης των συνδέσμων, διαπιστώνεται εύκολα ότι είναι κρίσιμο ζήτημα αφού ο αριθμός των δυνατών συνδέσμων, αν περιοριστούμε σε δυαδικές μεταβλητές με τιμή 1, είναι της τάξης $O(m2^{m-1})$. Ωστόσο η μορφή της συνάρτησης μέτρου επιτρέπει την μείωση του μέσου υπολογιστικού χρόνου σε λογικά επίπεδα. Ας σημειωθεί ότι αν $p(A=1) \leq p_s$ ή $p(B=1) \leq p_s$ τότε είναι φανερό ότι θα είναι και $p(A=1, B=1) \leq p_s$. Αξιοποιώντας αυτή την παρατήρηση, μπορεί αρχικά να γίνει αναζήτηση των μεμονωμένων στοιχείων (π.χ. A=1) κατά στήλη που εμφανίζουν συχνότητα (στήριξη) μεγαλύτερη του κατωφλιού p_s. Αυτό επιτυγχάνεται με μια γραμμική σάρωση όλου του πίνακα. Ένα στοιχείο (ή ένα σύνολο στοιχείων) που η συχνότητα εμφάνισής του στο πίνακα δεδομένων ξεπερνά το κατώφλι p_s ονομάζεται «συχνό». Μπορούμε λοιπόν να θεωρήσουμε όλα τα δυνατά ζευγάρια των συχνών στοιχείων τάξεως 1 ως υποψήφια συχνά στοιχεία τάξεως 2. Η διαπίστωση του ποια από αυτά τα σύνολα στοιχείων τάξεως 2 είναι πράγματι συχνά γίνεται με νέα σάρωση του πίνακα δεδομένων.

Γενικότερα, όταν μεταβαίνουμε από συχνά σύνολα τάξεως k-1 σε συχνά σύνολα τάξεως k, μπορούμε να εξαιρέσουμε («κλαδέψουμε» - prune) κάθε σύνολο

τάξεως k που περιλαμβάνει ένα υποσύνολο k-1 στοιχείων, τα οποία δεν συνιστούν συχνό σύνολο στο k-1 επίπεδο. Ας σημειωθεί ότι διαδικασία αυτή της εξαίρεσης, χωρίς να απαιτεί νέα σάρωση των δεδομένων, μειώνει σημαντικά τον υπολογιστικό φόρτο στη συνέχεια. Η διαδικασία που ακολουθείται σύμφωνα με τον αλγόριθμο a priori σκιαγραφείται στη συνέχεια μένα απλό παράδειγμα.

Υποθέτουμε ένα πίνακα δεδομένων με πέντε στήλες (m=5) που αντιστοιχούν σε ισάριθμες δυαδικές μεταβλητές Α, Β, Γ, Δ και Ε. Μπορούμε επίσης να θεωρήσουμε ότι κάθε μεταβλητή αντιπροσωπεύει ένα διακεκριμένο προϊόν και ότι κάθε γραμμή του πίνακα δεδομένων αντιστοιχεί σε ένα καλάθι αγορών (μια εμπορική συναλλαγή). Έτσι μια γραμμή του πίνακα της μορφής (1, 0, 0, 1, 0) ερμηνεύεται ως ένα καλάθι αγορών, το οποίο περιλαμβάνει μόνο τα προϊόντα Α και Δ. Το πλήθος των γραμμών του πίνακα, αν και παίζει καθοριστικό ρόλο στον υπολογισμό του υπολογιστικού φόρτου, δεν ενδιαφέρει για την παρουσίαση του αλγόριθμου. Θεωρούμε επίσης τα κατώφλια στήριξης και ακρίβειας p_s και p_a αντίστοιχα.

1. Όλα στοιχεία (σύνολα τάξεως 1, δηλαδή μονοσύνολα) {Α=1}, {Β=1}, {Γ=1}, {Δ=1} και {Ε=1} είναι αρχικά υποψήφια για χαρακτηρισμό ως συχνά.

2. Γίνεται μια σάρωση ($1^{η}$) του πίνακα δεδομένων για να βρεθούν τα στοιχεία που είναι πράγματι συχνά, δηλαδή εκείνα των οποίων η συχνότητα εμφάνισης (στήριξη) ξεπερνά το κατώφλι p_s. Έστω ότι $p(A=1) > p_s$, $p(B=1) > p_s$, $p(Γ=1) > p_s$, $p(Δ=1) < p_s$ και $p(E=1) < p_s$. Η λίστα λοιπόν με τα συχνά στοιχεία τάξης 1 είναι ακόλουθη: {Α=1}, {Β=1}, {Γ=1}

Υποψήφια συχνά στοιχεία τάξης 1	Συχνά στοιχεία τάξης 1	
{Α=1} {Β=1} {Γ=1} {Δ=1} {Ε=1}	{Α=1} {Β=1} {Γ=1}	Διαδικασία : Σάρωση του πίνακα δεδομένων

3. Όλα τα δυνατά ζευγάρια των συχνών στοιχείων τάξης 1 είναι υποψήφια για χαρακτηρισμό ως συχνά στοιχεία τάξης 2. Γίνεται σάρωση ($2^{η}$) του πίνακα δεδομένων για να βρεθούν τα στοιχεία τάξης 2 που είναι πράγματι συχνά, η οποία δίδει τα αποτελέσματα του ακόλουθου πίνακα:

Υποψήφια συχνά στοιχεία τάξης 2	Συχνά στοιχεία τάξης 2	
{A=1, B=1} {A=1, Γ=1} {B=1, Γ=1}	{A=1, B=1} {B=1, Γ=1}	Διαδικασία : Σάρωση του πίνακα δεδομένων

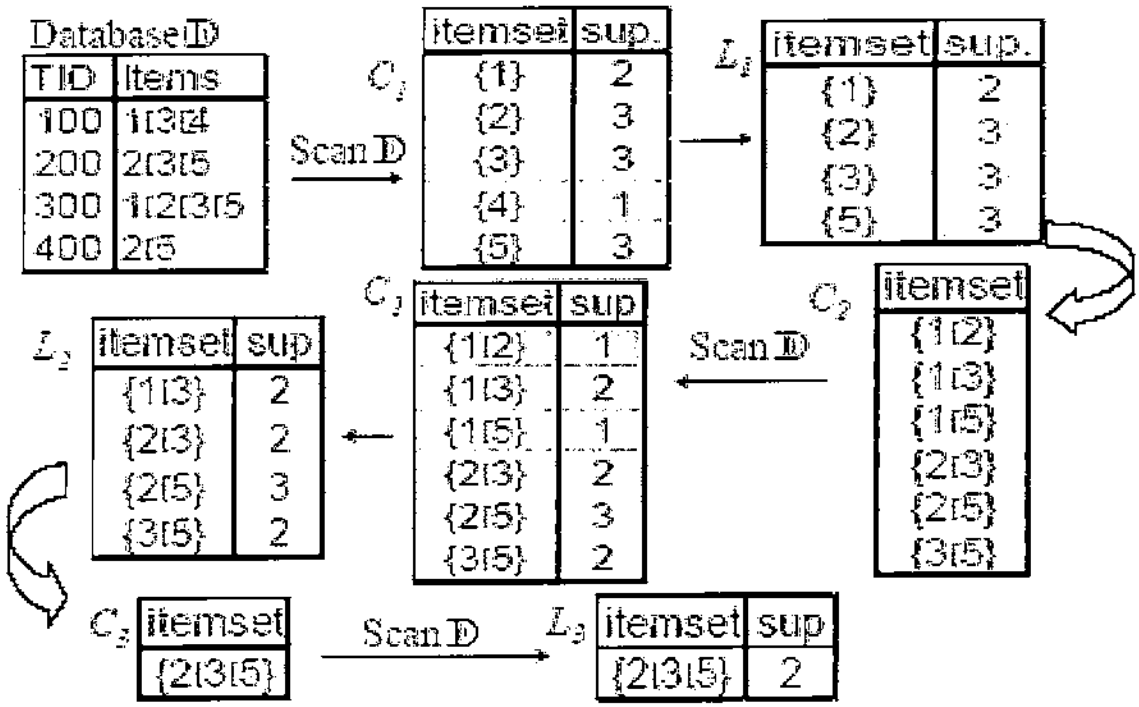
4. Υποψήφιο συχνό στοιχείο τάξης 3 είναι το σύνολο {A=1, B=1, Γ=1} που προέρχεται από συνδυασμό των συχνών στοιχείων τάξης 2. Για την διαπίστωση όμως τώρα αν το σύνολο αυτό είναι πράγματι συχνό ή όχι δεν χρειάζεται άλλη σάρωση του πίνακα δεδομένων. Αυτό γίνεται με κλάδεμα της λίστας. Πράγματι αφού το σύνολο {A=1, B=1, Γ=1} περιέχει το υποσύνολο {A=1, Γ=1}, το οποίο δεν είναι συχνό τάξης 2, δεν μπορεί να είναι και το ίδιο συχνό και εξαιρείται από τη λίστα. Παύει στο σημείο αυτό η διαδικασία αναζήτησης λιστών ανώτερης τάξης.

Υποψήφια συχνά στοιχεία τάξης 3	Συχνά στοιχεία τάξης 3	Διαδικασία : Εξαίρεση (κλάδεμα)
{A=1, B=1, Γ=1}	∅	

5. Σε ένα τελευταίο στάδιο, γίνεται έλεγχος της ακρίβειας των συνδέσμων στις λίστες των συχνών στοιχείων με βάση το κατώφλι ακρίβειας p_a, για να εξαχθούν οι σύνδεσμοι υπό μορφή κανόνων.

Σχετικά με την πολυπλοκότητα του αλγόριθμου, είναι φανερό ότι στη χειρότερη περίπτωση που όλα τα δυνατά σύνολα αποδεικνύονται συχνά, ο αλγόριθμος είναι εκθετικού χρόνου. Όμως στην πράξη επειδή οι πίνακες, δεδομένων στους οποίους κυρίως εφαρμόζεται ο αλγόριθμος a priori χαρακτηρίζονται από σποραδικότητα, η

μεγαλύτερη τάξη συχνών στοιχείων (πληθάριθμος των συνόλων στη λίστα τελευταίας τάξης) είναι πολύ μικρή σε σχέση με το m (πλήθος μεταβλητών). Τούτο ενισχύεται επιπλέον όταν το κατώφλι στήριξης είναι αρκετά μεγάλο.



6. Προχωρούμε έτσι όσο θέλουμε (ή μέχρι τα σύνολα να είναι κενά).

Σχήμα 20. Γραφικό παράδειγμα Εκτέλεσης του Apriori

Για να δούμε πώς μας βοηθά ο a-priori αλγόριθμος ας δούμε το ακόλουθο παράδειγμα:

Έστω η συσχέτιση Baskets(BID,item) με 10^8 εγγραφές, 10^7 καλάθια των 10 αντικειμένων κι έστω 100,000 διαφορετικά αντικείμενα.

Η ερώτηση SQL για να βρούμε τα συχνά ζεύγη αντικειμένων θα ήταν:

```
Select b1.item, b2.item, count(*)
from Baskets b1, Baskets b2
where b1.BID= b2.BID AND b1.item< b2.item
group by b1.item, b2.item
HAVING count(*)>s;
```

Στο join κάθε καλάθι συνεισφέρει $\binom{10}{2}$=45 ζεύγη. Έτσι, το join έχει $4,5*10^8$ εγγραφές. Μπορούμε να αντικαταστήσουμε το Basket(BID,item) με το query:

```
Select *
from Baskets
group by item
HAVING count(*)>=s;
```

Αν s=0.01, τότε το πολύ 1000 ομάδες αντικειμένων περνούν τη συνθήκη HAVING, γιατί υπάρχουν 108 εμφανίσεις αντικειμένων κι ένα αντικείμενο χρειάζεται $0,01*10^7=10^5$ αυτών για να εμφανίζεται σε 1% των καλαθιών. Αν και το 99% των αντικειμένων απορρίπτονται a-priori, δεν πρέπει να θεωρήσουμε ότι το αποτέλεσμα Baskets έχει μόνο 10^6 εγγραφές. Στην πραγματικότητα όλες οι εγγραφές ίσως είναι για όλα τα συχνά αντικείμενα. Οπωσδήποτε, στις πραγματικές καταστάσεις, η μείωση του Baskets είναι σημαντική και το μέγεθος του join μειώνεται ανάλογα με το τετράγωνο της μείωσης του Baskets.

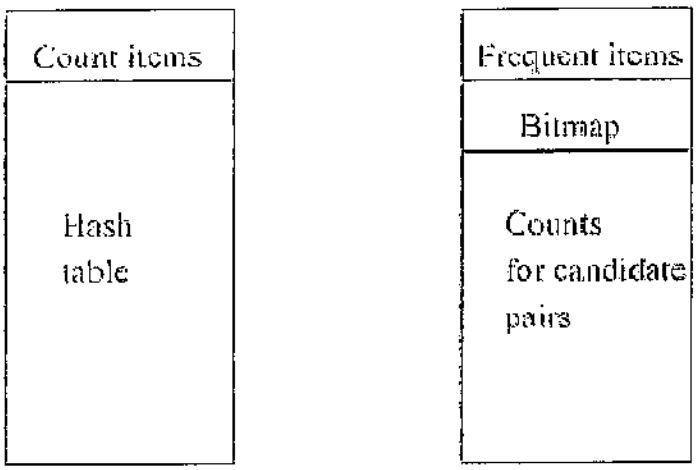
3.2 Βελτιώσεις του a-priori.

Εχουμε δύο τύπων βελτιώσεις:

1. Μείωση του μεγέθους των υποψηφίων συνόλων C_i, για i>=2.
2. Συνδυασμένες προσπάθειες για την εύρεση των L1,L2,L3,... σε ένα ή δύο περάσματα, από ένα πέρασμα για κάθε επίπεδο.

Αλγόριθμος PCY

Οι Park, Chen, Yu (1995) προτείνουν τη χρήση ενός πίνακα κατακερματισμού για να ορίσουν στο πρώτο πέρασμα ότι πολλά ζευγάρια δεν είναι συχνά. Βασίζεται στο γεγονός ότι η κύρια μνήμη είναι συνήθως μεγαλύτερη από τον αριθμό των αντικειμένων. Κατά τα δυο περάσματα για να βρούμε το L2 , η κύρια μνήμη είναι:



Έστω ότι τα δεδομένα βρίσκονται σε ένα αρχείο flat (ID, item1,..., itemn)

1. Πρώτο πέρασμα
    a. Μέτρα τις εμφανίσεις των αντικειμένων
    b. Για κάθε καλάθι αποτελούμενο από αντικείμενα {i1,...,ik} τοποθέτησε όλα τα ζεύγη σε έναν κάδο του πίνακα κατακερματισμού κι αύξησε το μετρητή του κάδου κατά ένα.
    c. Στο τέλος του περάσματος, καθόρισε το L1 (τα αντικείμενα που μετρήθηκαν τουλάχιστο s).
    d. Καθόρισε τους κάδους που μετρήθηκαν τουλάχιστον s. Ενα ζεύγος (i,j) δε μπορεί να είναι συχνό, εκτός αν τοποθετείται σε ένα συχνό κάδο, έτσι ζεύγη σε άλλους κάδους δε χρειάζεται να είναι υποψήφιαστοC2.

       Αντικατέστησε τον πίνακα κατακερματισμού με ένα bitmap με 1 bit για κάθε κάδο με τιμές 1 αν το bucket είναι συχνό και 0 αν όχι.
2. Δεύτερο πέρασμα
    a. Η κύρια μνήμη κρατά μια λίστα με τα συχνά αντικείμενα (δηλαδή το L1).
    b. Η κύρια μνήμη κρατά το bitmap συνοψίζοντας το αποτέλεσμα του κατακερματισμού από το πρώτο πέρασμα. Οι κάδοι πρέπει να χρησιμοποιούν 16 ή 32 bits για τη μέτρηση, αλλά συμπιέζονται στο 1 bit με τη χρήση του bitmap. Ετσι, ακόμη κι αν ο πίνακας κατακερματισμού καταλαμβάνει σχεδόν όλη τη μνήμη στο πρώτο

πέρασμα, το bitmap δεν καταλαμβάνει πάνω από το 1/16 της μνήμης στο δεύτερο πέρασμα.

c. Τέλος, η μνήμη κρατά έναν πίνακα με όλα τα υποψήφια ζεύγη. Ένα ζεύγος (i,j) είναι υποψήφιο αν ισχύουν όλα τα ακόλουθα:
   i. i ανήκει στο L1
   ii. j ανήκει στο L1
   iii. Το ζεύγος (i,j) τοποθετείται σε συχνό κάδο.

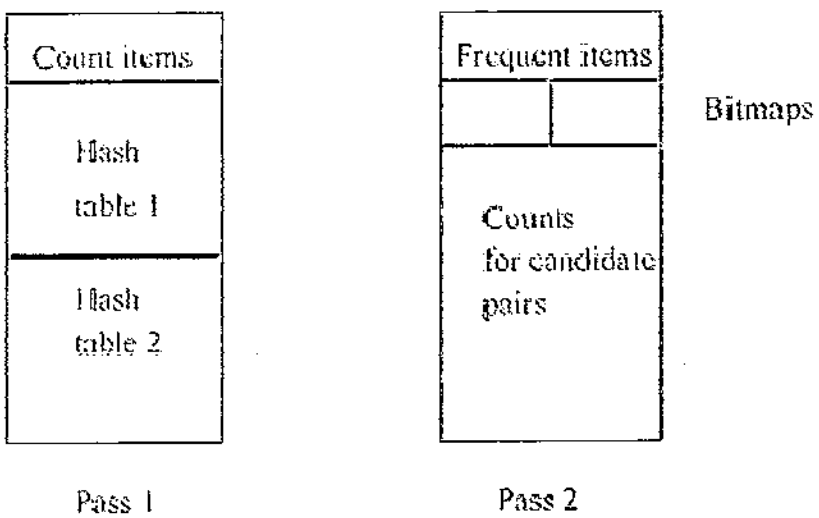
   Το (iii) είναι αυτό που διακρίνει τον PCY από τον a-priori

d. Κατά το δεύτερο πέρασμα 2, λαμβάνουμε υπόψιν κάθε καλάθι και κάθε ζεύγος των αντικειμένων κάνοντας τη δοκιμή που περιγράφεται προηγούμενα. Αν ένα ζεύγος ικανοποιεί τις 3 συνθήκες, πρόσθεσε τις φορές που εμφανίζεται στη μνήμη ή φτιάξε μια είσοδο γι' αυτό, αν δεν υπάρχει ήδη.

Η επέκταση του PCY "Iceberg"

1. Πολλαπλοί Πίνακες Κατακερματισμού (Fang και άλλοι, 1998)

Η μνήμη μοιράζεται σε δύο οι περισσότερους πίνακες κατακερματισμού στο πρώτο πέρασμα, όπως φαίνεται από το ακόλουθο σχήμα.

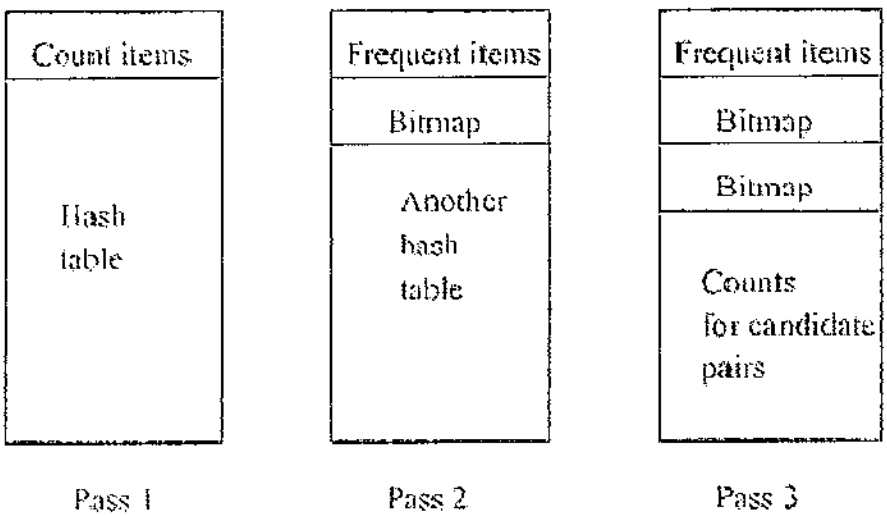


Στο δεύτερο πέρασμα, ένα bitmap αποθηκεύεται για κάθε πίνακα κατακερματισμού. Ο χώρος που απαιτείται για όλα τα bitmaps είναι ακριβώς ο ίδιος με αυτόν που χρειάζεται για ένα bitmap στον PCY, αφού ο συνολικός αριθμός των κάδων είναι ο ίδιος. Για να είναι ένα ζεύγος υποψήφιο στο C2 πρέπει:

a. να αποτελείται από αντικείμενα που ανήκουν στο L1
b. να τοποθετείται σε συχνό κάδο σε κάθε πίνακα κατακερματισμού.

2. Επαναλαμβανόμενοι πίνακες κατακερματισμού.

Αντί να ελέγχουμε για υποψήφια ζεύγη στο δεύτερο πέρασμα, εκτελούμε μια άλλη συνάρτηση κατακερματισμού, όχι για όλα τα ζεύγη αλλά μόνον για όσα περνούν το test του PCY. Δηλαδή, τα ζεύγη να αποτελούνται από αντικείμενα του L1 και στο πρώτο πέρασμα να έχουν τοποθετηθεί σε ένα συχνό κάδο.



Στο τρίτο πέρασμα, φυλάμε τα bitmaps κι από τους δύο πίνακες κατακερματισμού και ένα ζεύγος θα είναι υποψήφιο μόνον όταν:

a. και τα δύο αντικείμενα ανήκουν στο L1
b. Το ζεύγος τοποθετείται σε συχνό κάδο στο πρώτο πέρασμα
c. Το ζεύγος τοποθετείται, επίσης, σε συχνό κάδο στο δεύτερο πέρασμα

Εύρεση των συχνών συνόλων αντικειμένων σε δύο περάσματα

Οι προηγούμενες μέθοδοι προτιμώνται όταν θέλουμε συχνά ζεύγη αντικειμένων. Αν θέλουμε όλα τα συχνά σύνολα αντικειμένων, όσο μεγάλα κι αν είναι, χρειαζόμαστε περισσότερα περάσματα. Υπάρχουν αρκετές προσεγγίσεις για την εύρεση συχνών συνόλων αντικειμένων σε δύο το πολύ περάσματα.

1. Απλή προσέγγιση: Παίρνουμε ένα δείγμα δεδομένων όσο και η κύρια μνήμη. Εκτελούμε έναν κλιμακωτό αλγόριθμο (π.χ. A-priori) στην κύρια μνήμη κι έτσι δεν έχουμε κόστος I/O. Ελπίζουμε ότι το δείγμα δίνει τα πραγματικά συχνά σύνολα αντικειμένων.
    - Να σημειωθεί ότι πρέπει να μειωθεί το κατώφλι s. Π.χ. αν δείγμα μας είναι το 1% των δεδομένων πρέπει να χρησιμοποιηθεί κατώφλι s/100.
    - Μπορούμε να κάνουμε ένα πλήρες πέρασμα στα δεδομένα, για να επαληθεύσουμε ότι τα συχνά σύνολα αντικειμένων του παραδείγματος είναι πράγματι συχνά, αλλά ίσως χαθεί κάποιο σύνολο αντικειμένων που είναι συχνό για όλα τα δεδομένα, αλλά όχι στο δείγμα μας.
    - Για να ελαχιστοποιήσουμε τα λάθη, μπορούμε να μειώσουμε ακόμη λίγο το κατώφλι s για το δείγμα μας, κι έτσι να βρούμε περισσότερα υποψήφια συχνά σύνολα αντικειμένων για το πλήρες πέρασμα στα δεδομένα. Το πρόβλημα είναι ότι θα έχουμε πάρα πολλά υποψήφια συχνά σύνολα αντικειμένων για να χωρέσουν στην κύρια μνήμη.
2. SON95 (Savasere και άλλοι, 1995).

    Διάβασε τα υποσύνολα των δεδομένων στην κυρία μνήμη κι εφάρμοσε την απλή προσέγγιση για την εύρεση υποψηφίων συνόλων. Κάθε καλάθι είναι μέρος ενός τέτοιου υποσυνόλου. Στο δεύτερο πέρασμα, ένα σύνολο είναι υποψήφιο αν ήταν υποψήφιο σε κάποιο ή περισσότερα υποσύνολα.
    Το σημείο-κλειδί είναι ότι ένα σύνολο αντικειμένων δε μπορεί να είναι συχνό για τα δεδομένα αν δεν είναι συχνό τουλάχιστον σε ένα υποσύνολό τους

3. Ο αλγόριθμος Toivonen (Toivonen, 1996)
    a. Πάρε ένα δείγμα στο μέγεθος της μνήμης. Εφάρμοσε την απλή προσέγγιση, αλλά με κατώφλι χαμηλότερο για να μη χαθούν συχνά

σύνολα. (π.χ. αν το δείγμα είναι 1% των δεδομένων, το κατώφλι να είναι s/125).

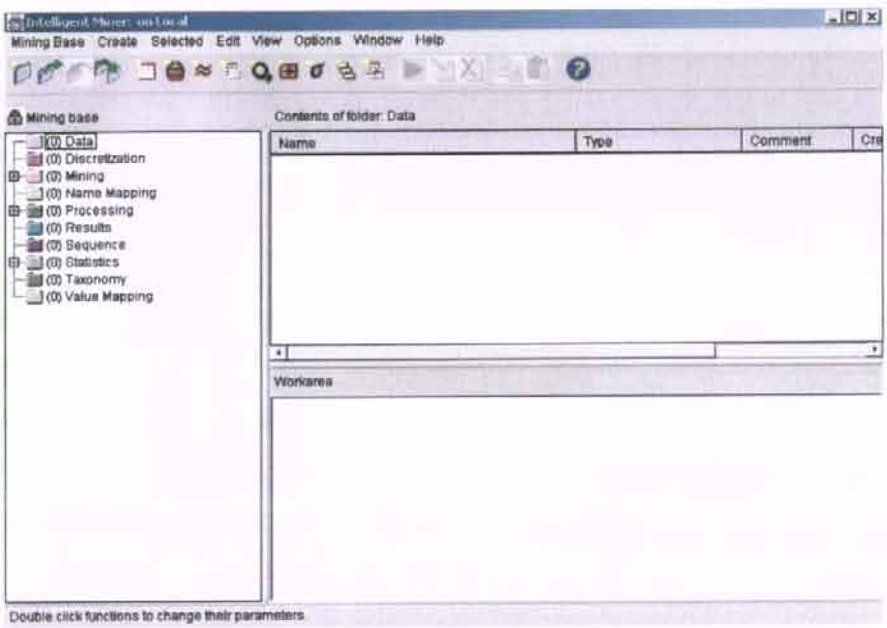
b. Πρόσθεσε στα υποψήφια του δείγματος το "αρνητικό όριο". Ως αρνητικό όριο ορίζεται το σύνολο S των αντικειμένων που δεν είναι συχνό στο δείγμα, αλλά όλα τα άμεσα υποσύνολά του είναι συχνά.

c. Μέτρα όλα τα υποψήφια σύνολα αντικειμένων και το αρνητικό όριο. Αν κανένα μέλος του αρνητικού ορίου δεν είναι συχνό για όλα τα δεδομένα, τότε τα συχνά σύνολα αντικειμένων είναι τα υποψήφια που υπολογίστηκαν με το κατώφλι.

d. Δυστυχώς, αν ένα μέλος του αρνητικού ορίου είναι συχνό, τότε δεν ξέρουμε κάποιο από τα υπερσύνολά του είναι επίσης, συχνό. Έτσι, η όλη διαδικασία πρέπει να επαναληφθεί (ή να δεχθούμε ότι το αποτέλεσμα είναι ικανοποιητικό παρά τα λίγα σφάλματα που υπάρχουν).

4. DB2 INTELLIGENT MINER FOR DATA

Ο Intelligent Miner for Data είναι μία ανεξάρτητη από την DB2 εφαρμογή πού έχει την ικανότητα να αναλύει κάποιες μεθόδους εξόρυξης δεδομένων (http://www-306.ibm.com/software/data/iminer/fordata/). Είναι ένα καθαρά γραφικό περιβάλλον. Αυτό σημαίνει ότι τα αποτελέσματα που παράγει απεικονίζονται μόνο στην οθόνη και δεν μπορούν να δεχθούν καμία επεξεργασία. Υπάρχει βέβαια η δυνατότητα να εξαχθούν σε flat αρχεία (αρχεία κειμένου) και από εκεί να χρησιμοποιηθούν από τις διάφορες γλώσσες προγραμματισμού για επεξεργασία. Η αρχική εικόνα του Intelligent Miner φαίνεται στο σχήμα 21:



Σχήμα 21 : αρχική εικόνα του Intelligent Miner.

Στο αριστερό μέρος της αρχικής οθόνης (Σχήμα 21) υπάρχουν κάποιοι κατάλογοι οι οποίοι αντιπροσωπεύουν κάποια αντικείμενα και συναρτήσεις τα οποία μπορούν να δημιουργηθούν. Κάθε αντικείμενο και συνάρτηση έχει κάποιες παραμέτρους (η κάθε μία διαφορετικές). Με τον Intelligent Miner μπορούν να δημιουργηθούν αντικείμενα και συναρτήσεις και να αποθηκευτούν στον δίσκο μαζί με τα «μεταδεδομένα» τους (τις ιδιότητές τους). Τα αντικείμενα που χρησιμοποιούνται είναι τα εξής:

Data : το όνομα και η τοποθεσία των δεδομένων εισόδου για να χρησιμοποιηθούν από τις συναρτήσεις.

Descretization : κατανομή των εγγραφών σε αναγνωρίσιμες ομάδες.

Name mapping : προσδιορισμός των τιμών που αντιστοιχούν σε πεδία τύπου Categorical.

Results : το όνομα και η τοποθεσία των αποτελεσμάτων που έχουν δημιουργηθεί από τις συναρτήσεις.

Taxonomy : ιεραρχίες συσχετίσεων μεταξύ διαφορετικών κατηγοριών ενός στοιχείου.

Value Mapping : ο προσδιορισμός των τιμών που αντιστοιχούν σε άλλες τιμές.

Οι συναρτήσεις στον Intelligent Miner είναι οι εξής:

Mining : εδώ δηλώνονται οι παράμετροι για κάθε συνάρτηση εξόρυξης δεδομένων. Αυτή θα είναι και η συνάρτηση με την οποία θα εργαστούμε στο επόμενο κεφάλαιο.

Preprocessing : δήλωση των παραμέτρων των μοναδικών προ-επεξεργαστικών διαδικασιών.

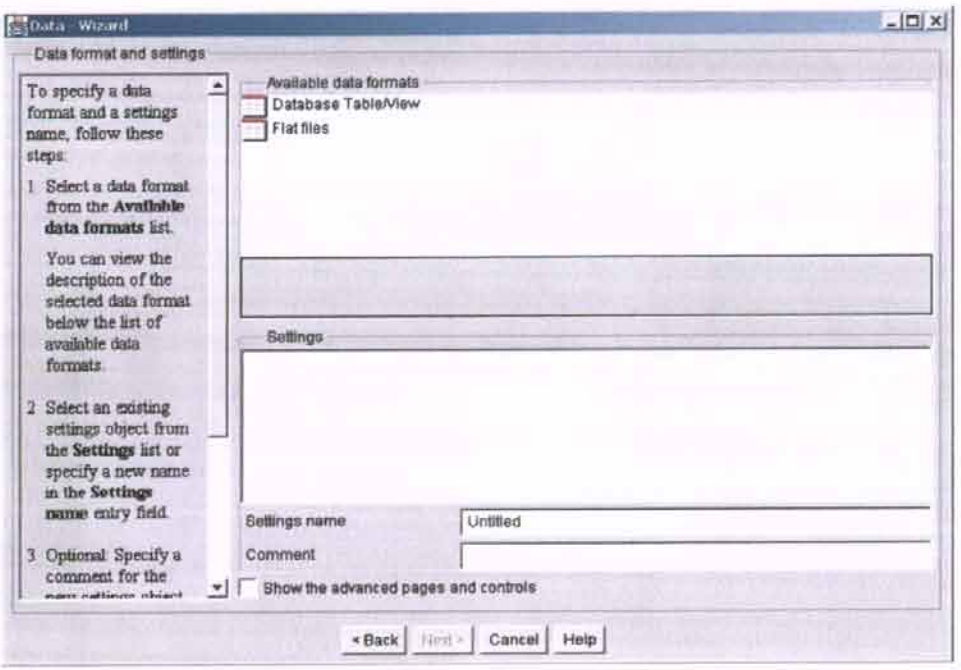
Sequence : Ο προσδιορισμός αρκετών συναρτήσεων οι οποίες μπορούν να ξεκινήσουν με καθορισμένη συχνότητα.

Statistics : δηλώνονται οι παράμετροι για κάθε στατιστική συνάρτηση.

Φυσικά όλες οι παραπάνω δυνατότητες δεν χρειάζονται για τις προσεγγίσεις που χρησιμοποιούμε. Τα μόνα αντικείμενα που θα χρησιμοποιήσουμε είναι το Data και το Results και από τις συναρτήσεις μόνο την Mining.

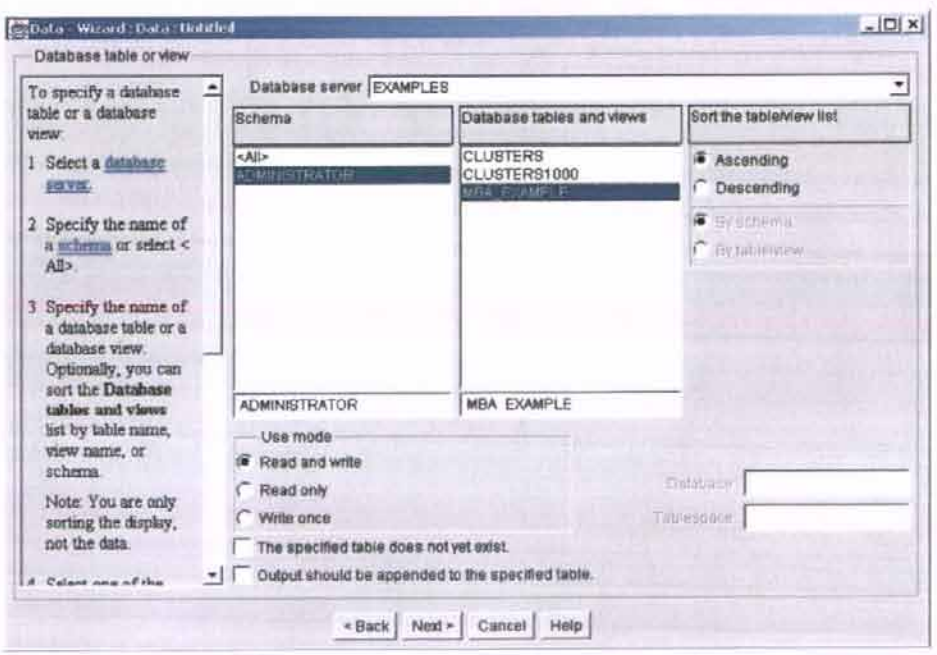
Κάνοντας δεξί κλικ στο αντικείμενο Data εμφανίζεται μία επιλογή με την οποία δηλώνουμε τα δεδομένα εισόδου. Με την επιλογή αυτή εμφανίζεται ένας οδηγός. Το πρώτο παράθυρου του οδηγού δίνουμε ένα όνομα στα δεδομένα και επιλέγουμε την πηγή τους, δηλαδή εάν θα προέρχονται από μία σχεσιακή βάση

δεδομένων ή αν θα προέρχονται από flat αρχείο Σχήμα 22 (αρχείο κειμένου τύπου ASCII). Στις δικές μας προσεγγίσεις χρησιμοποιήσαμε μόνο δεδομένα εισόδου τύπου πινάκων.



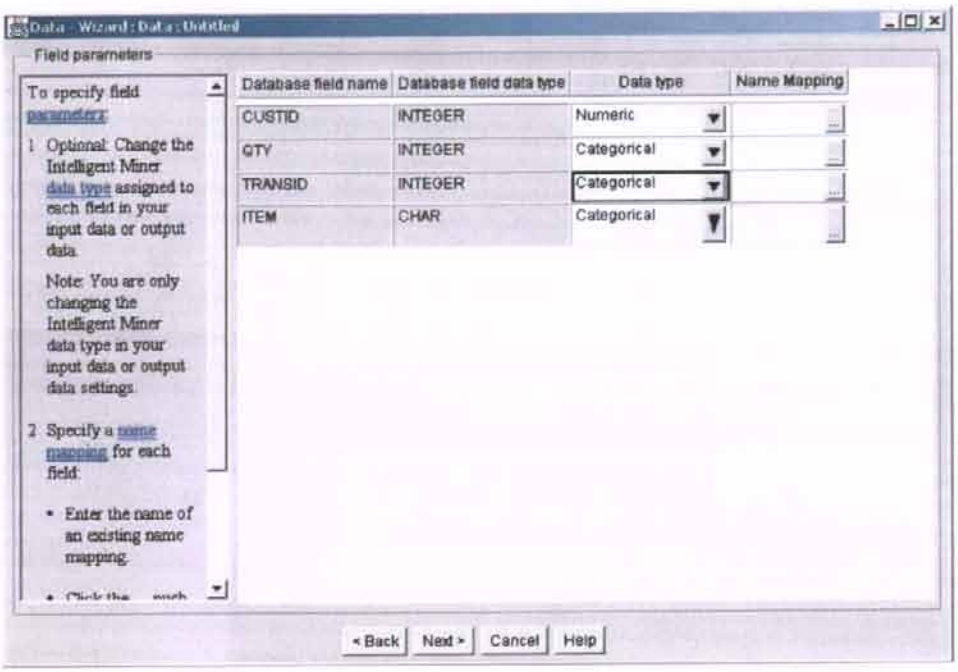
Σχήμα 22: παράθυρο οδηγού για το αντικείμενο Data.

Στο επόμενο βήμα του οδηγού επιλέγουμε τον πίνακα στον οποίο υπάρχουν τα δεδομένα εισόδου Σχήμα 23



Σχήμα 23: επιλογή πίνακα για τα δεδομένα εισόδου.

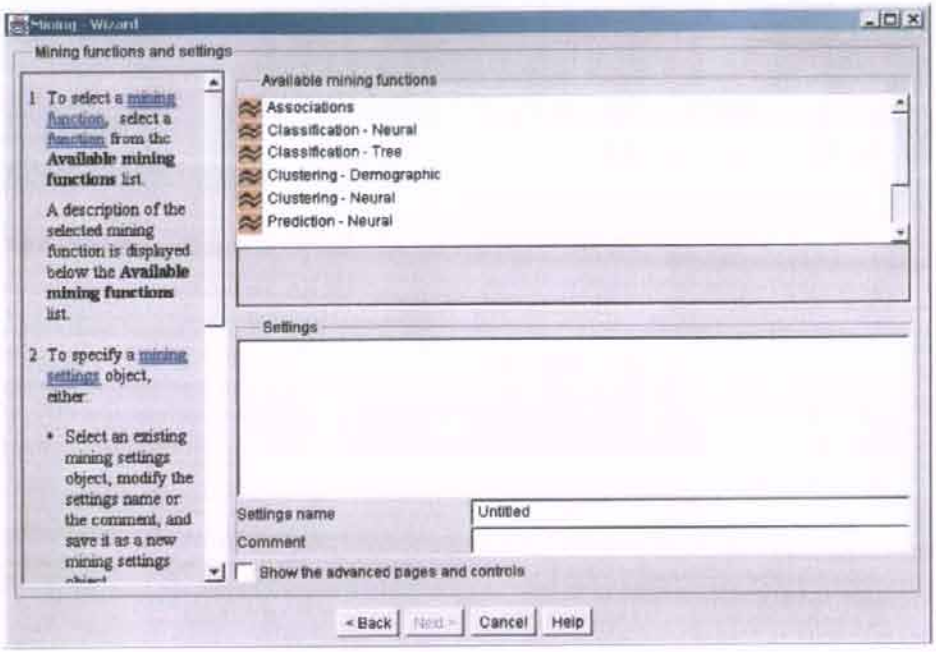
Έπειτα επιλέγουμε τον τύπο των δεδομένων για το κάθε πεδίο του πίνακα (βλέπε Σχήμα 24).



Σχήμα 24: τύποι δεδομένων για το κάθε πεδίο του πίνακα.

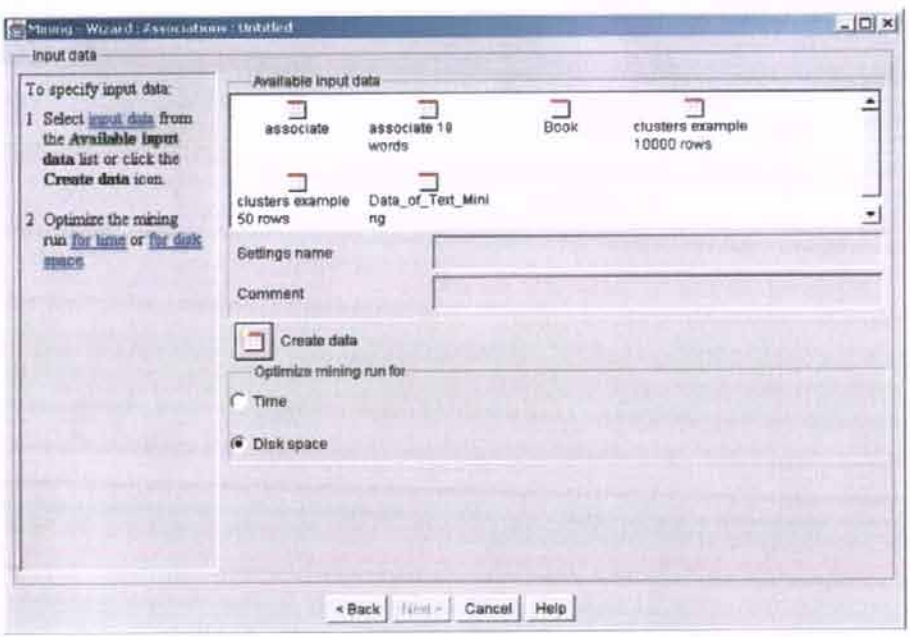
Δύο τύποι θα χρησιμοποιηθούν για όλες τις αναλύσεις: Continuous για τα πεδία που είναι αριθμοί και θα αντιμετωπιστούν από τον αλγόριθμο ως αριθμοί και Categorical που είναι πεδία τύπου κειμένου ή αριθμών αλλά συμπεριφέρονται σαν απλά σύμβολα.

Για να τρέξουμε οποιοδήποτε αλγόριθμο εξόρυξης δεδομένων χρησιμοποιούμε την συνάρτηση Mining. Πατώντας δεξί κλικ εμφανίζεται η επιλογή καινούργιας συνάρτησης. Το πρώτο πράγμα που πρέπει να κάνουμε είναι να επιλέξουμε τον αλγόριθμο που θα χρησιμοποιήσουμε από την λίστα του Σχήματος 25.



Σχήμα 25: επιλογή αλγορίθμου.

Στην συνέχεια και αφού έχουμε δώσει ένα όνομα στην συνάρτηση εμφανίζεται το παράθυρο του Σχήματος 26 στο οποίο δηλώνουμε τα δεδομένα εισόδου που ήδη δημιουργήσαμε.



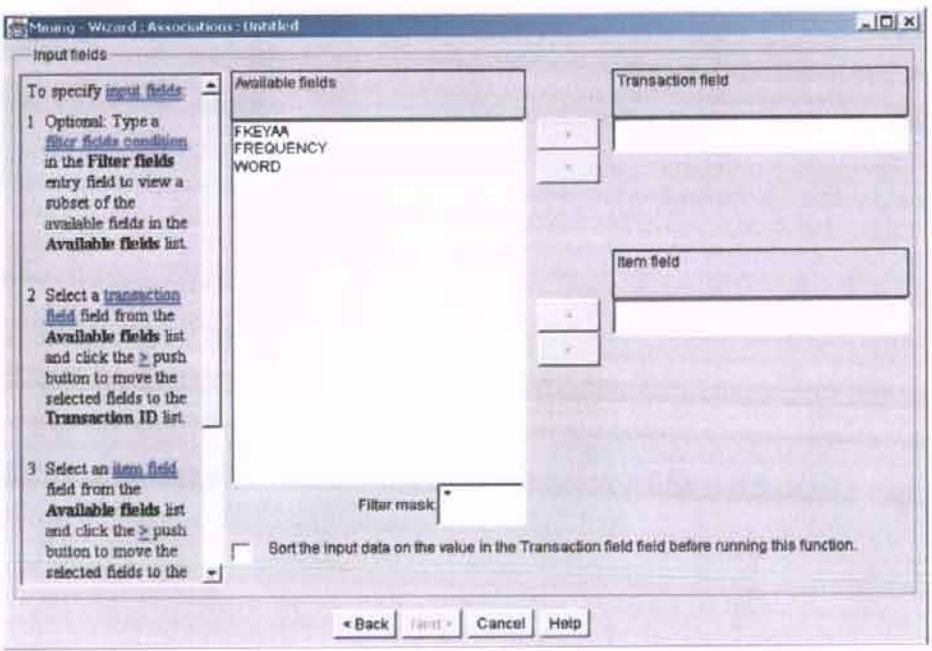
Σχήμα 26: επιλογή των δεδομένων εισόδου.

Στο παραπάνω σχήμα υπάρχει και μία ακόμη παράμετρος: αυτή της βελτιστοποίησης του αλγορίθμου ανάλογα με τον χρόνο ή τον χώρο στον δίσκο. Η επιλογή του χρόνου (time) σημαίνει ότι τα πρόσφατα δεδομένα του αλγορίθμου τοποθετούνται στην κύρια μνήμη του υπολογιστή ενώ με την τιμή Disk space ο

αλγόριθμος δουλεύει στον δίσκο και είναι πολύ πιο αργός. Βέβαια όταν τα δεδομένα είναι πάρα πολλά η δεύτερη επιλογή είναι σχεδόν αναγκαία διότι ο χώρος της κύριας μνήμης δεν φτάνει για την υλοποίηση του αλγορίθμου.

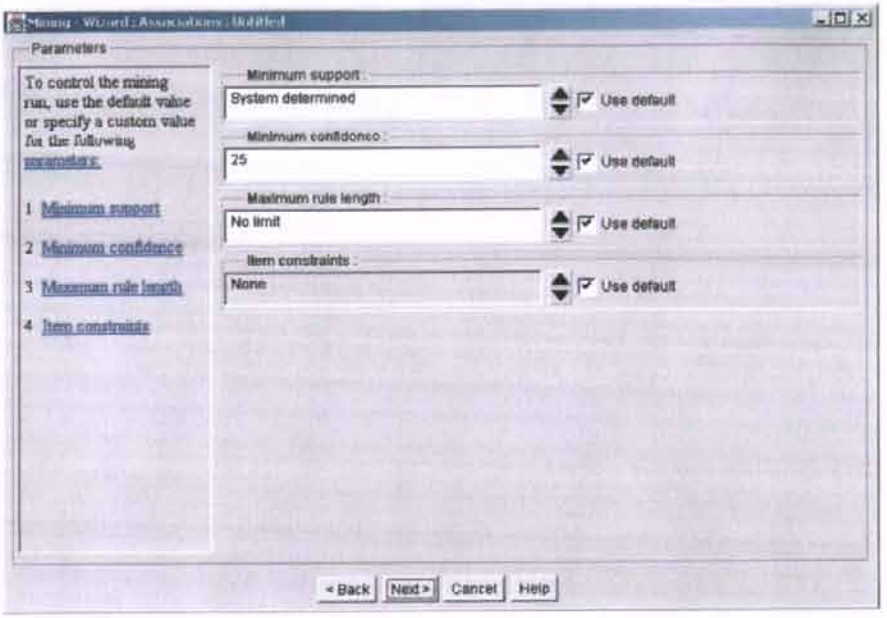
Από εδώ και πέρα οι ρυθμίσεις του κάθε αλγορίθμου διαφέρουν. Παρακάτω θα εξηγηθούν οι ρυθμίσεις που γίνονται για τους αλγορίθμους «κανόνες συσχέτισης».

Στους κανόνες συσχέτισης πρέπει να δηλωθούν τα πεδία συναλλαγών και τα πεδία στοιχείων. Σχήμα 27.



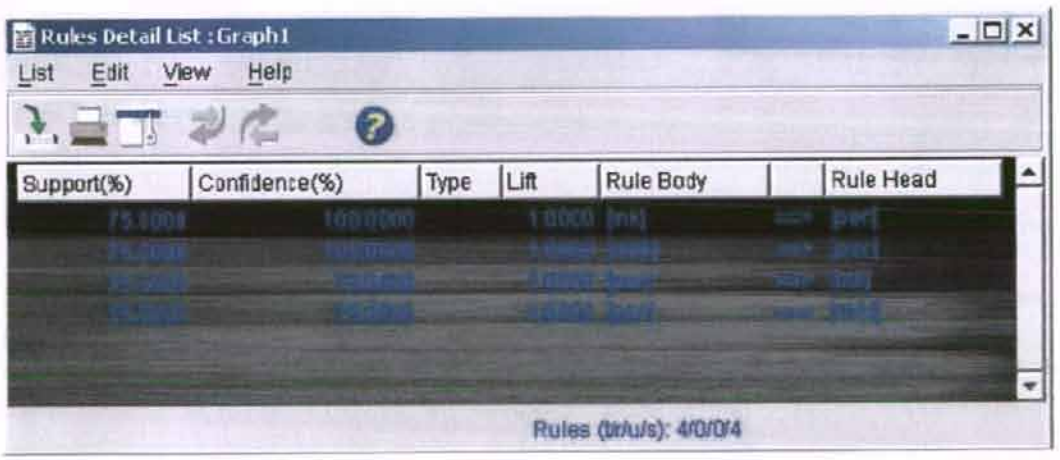
Σχήμα 27: πεδίο συναλλαγών και πεδίο στοιχείων.

Στην συνέχεια δηλώνονται τα ελάχιστα support και confidence (Σχήμα 28) όπως επίσης και πόσα στοιχεία το πολύ θα περιέχουν οι κανόνες συσχέτισης. Παράλληλα υπάρχει και η επιλογή item constraints η οποία καθορίζει ποιους κανόνες θα συμπεριλαμβάνονται ή δεν θα συμπεριλαμβάνονται στα αποτελέσματα.



Σχήμα 28: παράμετροι «κανόνων συσχέτισης».

Τέλος στην περίπτωση που γίνεται η χρήση των taxonomies δηλώνονται στο επόμενο στάδιο και έτσι ολοκληρώνεται ο οδηγός. Τα αποτελέσματα έχουν την εξής μορφή:



Σχήμα 29: αποτελέσματα αλγορίθμου «κανόνων συσχετίσεων»

Το αντικείμενο Results δείχνει τα αποτελέσματα όλων των αλγορίθμων σε γραφικό περιβάλλον.

5. ΕΦΑΡΜΟΓΗ ΤΩΝ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ ΣΤΙΣ ΗΛΕΚΤΡΟΝΙΚΕΣ ΚΑΤΑΘΕΣΕΙΣ

Σε αυτή την ενότητα περιγράφεται ο προσδιορισμός της σχέσης μεταξύ διαφορετικών τρόπων πληρωμών και καταθετών. Αυτές οι σχέσεις είναι μεγάλης σημασίας για την τράπεζα, διότι ενισχύουν την πώληση των προϊόντων με το να καταστήσουν την τράπεζα πιο ενήμερη για την συμπεριφορά των πελατών της και τον τρόπο πληρωμής που προτιμούν. Συμβάλλουν επίσης στη βελτίωση των υπηρεσιών της.

Ονομαστικά χρησιμοποιούνται οι ακόλουθοι τρόποι πληρωμών :

1.Μεταφορές κεφαλαίων μεταξύ των τραπεζών που συμμετέχουν στο σύστημα DIASTRANSFER.

2.Μεταφορές κεφαλαίων μεταξύ εθνικών και ξένων τραπεζών μέσω του συστήματος SWIFT.

3.Μεταφορές κεφαλαίων μέσα στην τράπεζα.

4.Μεταφορες κεφαλαίων μέσα στη τράπεζα στην τοκοφόρο ημερομηνία.

5.Διαταγές πληρωμής για τον Οργανισμό Τηλεπικοινωνιών Ελλάδας (ΟΤΕ) και για την Δημόσια Επιχείρηση Ηλεκτρισμού (ΔΕΗ).

6.Διαταγές πληρωμής για να καταβληθούν οι συνεισφορές των εργοδοτών (ΙΚΑ).

7.Διαταγές πληρωμής Φ.Π.Α.

8.Διαταγές πληρωμής με πιστωτική κάρτα.

Για τη συγκεκριμένη περίπτωση μελέτης χρησιμοποιήθηκε ένα τυχαίο δείγμα ηλεκτρονικών καταθετών, από τράπεζες, μεμονωμένα άτομα και επιχειρήσεις.

Μία δίτιμη τιμή ορίζεται σε κάθε διαφορετικό τύπο πληρωμής ανάλογα με το αν η πληρωμή έχει διευθυνθεί από τους χρήστες ή όχι. Σε περίπτωση που ο χρήστης είναι «επιχείρηση» λαμβάνει την τιμή 0 αλλιώς την τιμή 1.

Ένα παράδειγμα του ανωτέρου συνόλου στοιχείων παρουσιάζεται στον πίνακα 30.

User	Comp	Dias Transfer P.O.	Swift P.O.	Funds Transfer	Forward Funds Transfer	GTO PPC S.O.	SII P.O.	VAT P.O.	Cr. Card P.O.
...	...	...	...	...	...	...	...	...	...
User103	0	T	F	T	T	F	F	F	T
User104	1	T	T	T	T	F	T	T	F
User105	1	T	F	T	T	T	T	T	T
User106	0	T	F	T	F	T	F	F	T
...	...	...	...	...	...	...	...	...	...

Πίνακας 30. Διαφορετικοί τρόποι πληρωμής και καταθέτες

Προκειμένου να αναλυθούν οι κανόνες συσχέτισης, χρησιμοποιήθηκε ο αλγόριθμος Apriori και πιο συγκεκριμένα η υλοποίηση του από το εργαλείο DB2 INTELLIGENT MINER FOR DATA που περιγράφηκε στην προηγούμενη ενότητα. Αυτοί οι κανόνες αποτελούν μέρος της φόρμας «εάν έχουμε ένα γεγονός τότε αναμένεται κάποια συνέπεια». Στην περίπτωση αυτής της μελέτης που αναφέρθηκε προηγμένως ένας άλλος στόχος είναι η παραγωγή και η δοκιμή των προβλέψεων για τον όγκο των συναλλαγών των ηλεκτρονικών καταθετών σε σχέση με τους ενεργούς χρήστες. Οι οικονομικές συναλλαγές για όλες τις διαταγές πληρωμής ή της μόνιμες διαταγές που πραγματοποιεί ένας χρήστης αποκλείουν τις συναλλαγές σχετικά με το περιεχόμενο των πληροφοριών, όπως την ισορροπία απολογισμού, τις λεπτομερείς συναλλαγές απολογισμού.

Ο όρος «ενεργός» περιγράφει το χρήστη που χρησιμοποιεί την συγκεκριμένη περίοδο τις ηλεκτρονικές υπηρεσίες που προσφέρει η τράπεζα. Οι « ενεργοί » χρήστες είναι υποομάδα των χρηστών.

Μία ημέρα ορίζεται ως η χρονική μονάδα. Ο αριθμός των ενεργών χρηστών είναι η μεταβλητή προαγγέλων όσο ο όγκος των οικονομικών συναλλαγών υποτίθεται ότι παριστάνει την μεταβλητή απάντησης.

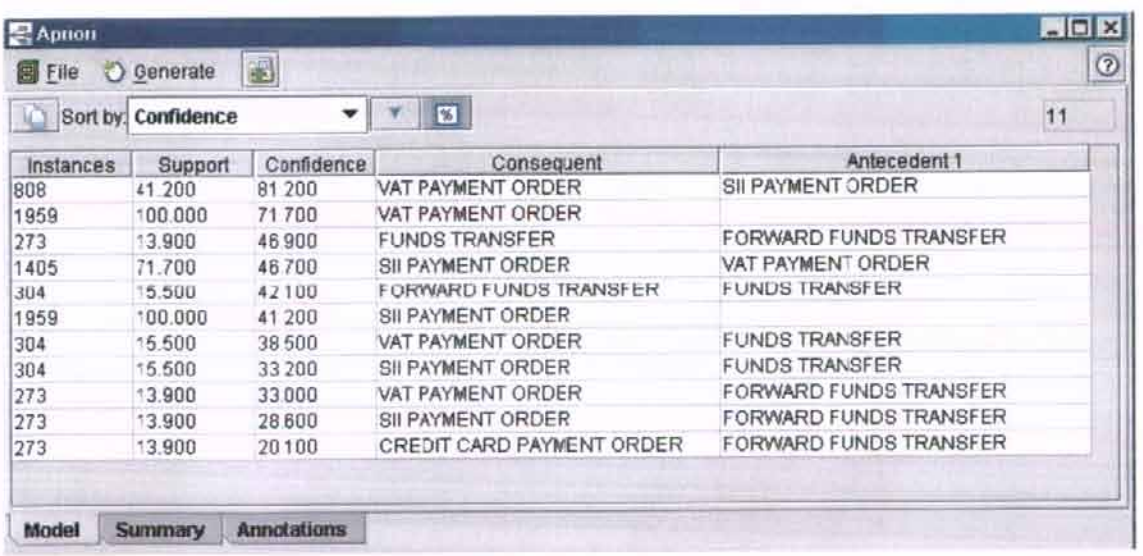
Ένα παράδειγμα σύμφωνα με τα παραπάνω είναι :

Transaction Day	Count_Of_Active_Users	Count_Of_Payments
...	...	...
27/8/2002	99	228
28/8/2002	107	385
29/8/2002	181	915
30/8/2002	215	859
...	...	...

Πίνακας 31. Ενεργοί χρήστες και ηλεκτρονικοί καταθέτες κατά την ημέρα συναλλαγής

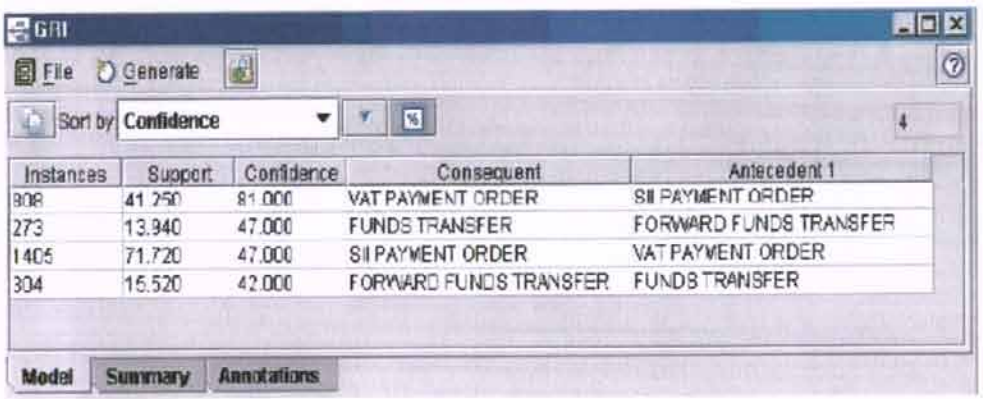
Η ημερομηνία με την οποία το σύνολο στοιχείων ισχύει μετρά από τις 20 Απριλίου του 2001 μέχρι τις 12 Δεκεμβρίου του 2002. Το σύνολο δεδομένων περιλαμβάνει μόνο το στοιχείο των ενεργών ημερών (διακοπές, Σαββατοκύριακα δεν περιλαμβάνονται) το οποίο συνεπάγεται 387 περιστατικά.

Πειραματικά αποτελέσματα



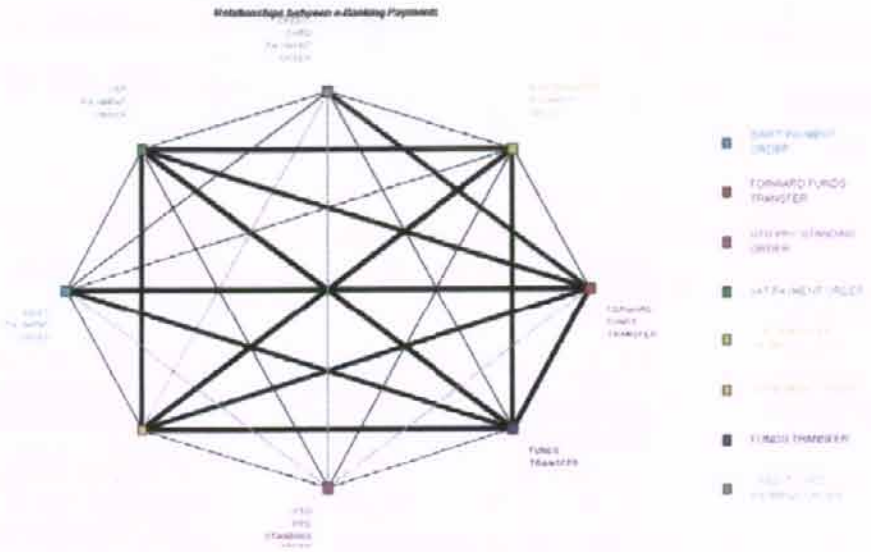
Σχήμα 32. Κανόνες συσχέτισης με βάση τον αλγόριθμο apriori.

Με τη χρήση της μεθόδου Apriori και θέτοντας κανόνα ελάχιστης υποστήριξης = 10 και ελάχιστο κανόνα εμπιστοσύνης = 20, καθορίστηκαν 11 κανόνες και παρουσιάστηκαν στο σχήμα.32 που βλέπουμε επάνω. Χρησιμοποιώντας με την μέθοδο της γενικευμένης επαγωγής τον κανόνα ελάχιστης υποστήριξης = 10 και τον κανόνα ελάχιστης υποστήριξης = 40 έχουμε σαν αποτέλεσμα τους 4 κανόνες του σχήματος 33.



Σχήμα 33. Κανόνες συσχέτισης με βάση τη μέθοδο γενικευμένης επαγωγής κανόνα.

Μετά την σύγκριση μεταξύ των κανόνων που λαμβάνονται με τις δύο μεθόδους μπορεί να συναχθεί το συμπέρασμα ότι ο ισχυρότερος είναι *AN* διαταγή πληρωμής ΙΚΑ *TOTE* διαταγή πληρωμής ΦΠΑ (confidence=81). (Rule 1). Ο κανόνας αυτός ισχύει και για την αντίστροφη περίπτωση. Δηλαδή : *AN* διαταγή πληρωμής ΦΠΑ *TOTE* διαταγή πληρωμής ΙΚΑ (confidence= 47). (Rule 2). Δύο άλλοι κανόνες με εμπιστοσύνη > 40 είναι οι ακόλουθοι : *AN Μεταφορά Κεφαλαίων TOTE* Μεταφορά Κεφαλαίων στην τοκοφόρο ημερομηνία (confidence=47), (Rule 3) – *AN* Μεταφορά Κεφαλαίων στην τοκοφόρο ημερομηνία *TOTE* Μεταφορά Κεφαλαίων (confidence=42), (Rule 4). Βλέπουμε ότι υπάρχει ισχυρή σχέση μεταξύ της διαταγής πληρωμής Φ.Π.Α. και της διαταγής πληρωμής Ι.Κ.Α. Αυτή η ισχυρή σχέση μαζί με άλλες απεικονίζεται στην παρακάτω γραφική παράσταση ιστού.



Σχήμα 34. Σχέση μεταξύ διαταγών πληρωμής Φ.Π.Α. και Ι.Κ.Α.

Strong Links

Links	Field 1	Field 2
656	VAT PAYMENT ORDER = "T"	SII PAYMENT ORDER = "T"
128	FUNDS TRANSFER = "T"	FORWARD FUNDS TRANSFER = "T"
117	FUNDS TRANSFER = "T"	VAT PAYMENT ORDER = "T"
101	FUNDS TRANSFER = "T"	SII PAYMENT ORDER = "T"
90	FORWARD FUNDS TRANSFER = "T"	VAT PAYMENT ORDER = "T"
78	FORWARD FUNDS TRANSFER = "T"	SII PAYMENT ORDER = "T"
55	FORWARD FUNDS TRANSFER = "T"	CREDIT CARD PAYMENT ORDER = "T"
47	SWIFT PAYMENT ORDER = "T"	FORWARD FUNDS TRANSFER = "T"
40	DIASTRANSFER PAYMENT ORDER = "T"	SII PAYMENT ORDER = "T"
37	DIASTRANSFER PAYMENT ORDER = "T"	FUNDS TRANSFER = "T"
37	SWIFT PAYMENT ORDER = "T"	FUNDS TRANSFER = "T"
36	DIASTRANSFER PAYMENT ORDER = "T"	VAT PAYMENT ORDER = "T"

Medium Links

Links	Field 1	Field 2
33	SWIFT PAYMENT ORDER = "T"	VAT PAYMENT ORDER = "T"
31	SWIFT PAYMENT ORDER = "T"	SII PAYMENT ORDER = "T"
27	FUNDS TRANSFER = "T"	CREDIT CARD PAYMENT ORDER = "T"
26	CREDIT CARD PAYMENT ORDER = "T"	VAT PAYMENT ORDER = "T"
25	DIASTRANSFER PAYMENT ORDER = "T"	FORWARD FUNDS TRANSFER = "T"
25	DIASTRANSFER PAYMENT ORDER = "T"	SWIFT PAYMENT ORDER = "T"
24	VAT PAYMENT ORDER = "T"	GTO PPC STANDING ORDER = "T"
17	SII PAYMENT ORDER = "T"	GTO PPC STANDING ORDER = "T"
17	FUNDS TRANSFER = "T"	GTO PPC STANDING ORDER = "T"
17	SWIFT PAYMENT ORDER = "T"	CREDIT CARD PAYMENT ORDER = "T"
17	DIASTRANSFER PAYMENT ORDER = "T"	CREDIT CARD PAYMENT ORDER = "T"
15	DIASTRANSFER PAYMENT ORDER = "T"	GTO PPC STANDING ORDER = "T"

Weak Links

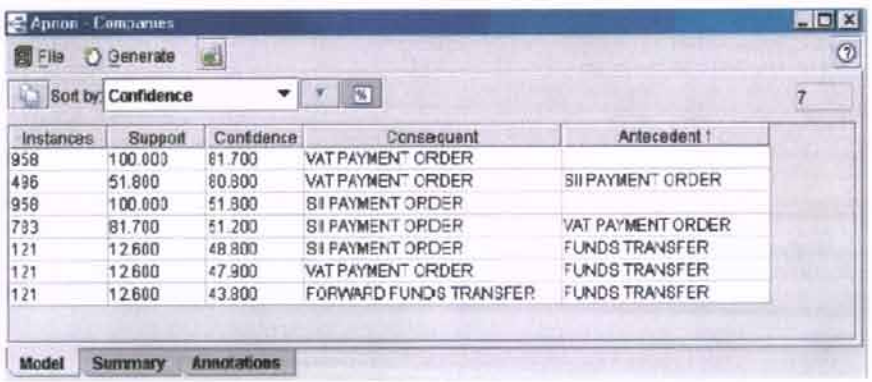
Links	Field 1	Field 2
13	CREDIT CARD PAYMENT ORDER = "T"	SII PAYMENT ORDER = "T"
10	CREDIT CARD PAYMENT ORDER = "T"	GTO PPC STANDING ORDER = "T"
7	FORWARD FUNDS TRANSFER = "T"	GTO PPC STANDING ORDER = "T"
6	SWIFT PAYMENT ORDER = "T"	GTO PPC STANDING ORDER = "T"

Σχήμα 35. Οι κανόνες συσχέτισης πρέπει να είναι εύκολα αντιληπτοί και χρήσιμοι.

Το ενδιαφέρον των ανωτέρω κανόνων είναι μεγάλο λαμβάνοντας υπόψη ότι γενικά ένας κανόνας αποκτά σημασία όταν γίνεται εύκολα αντιληπτός , απροσδόκητος

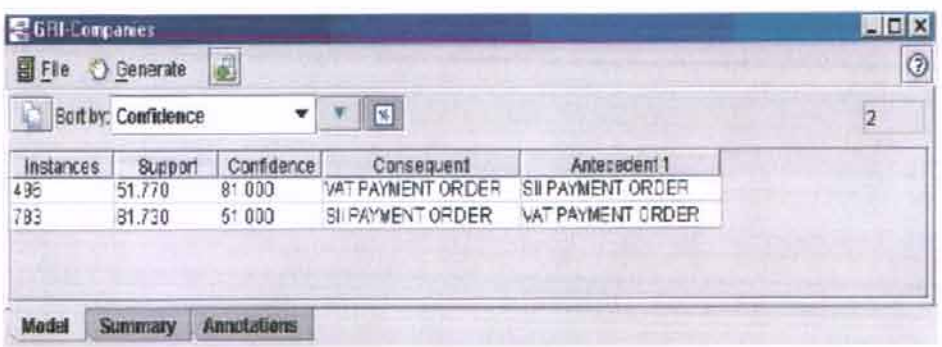
ενδεχομένως χρήσιμος και αγώγιμος ή να επικυρώνει κάποια υπόθεση που ένας χρήστης επιδιώκει να επιβεβαιώσει. Λαμβάνοντας υπόψη το γεγονός ότι οι τύποι πληρωμής που χρησιμοποιούνται στους κανόνες που εκθέτουν την υψηλότερη εμπιστοσύνη (διαταγή πληρωμής Φ.Π.Α., διαταγή πληρωμής Ι.Κ.Α.) είναι συνήθως πληρωμές που διευθύνονται από τις επιχειρήσεις . Μια μικρότερη υποδιαίρεση του δείγματος εξετάστηκε περιέχοντας μόνο τις επιχειρήσεις (958 χρήστες) για τις οποίες η μέθοδος Apriori και εφαρμόστηκε. Τα αποτελέσματα περιγράφονται παρακάτω.

Apriori : Χρησιμοποιώντας τη μέθοδο αυτή και θέτοντας ελάχιστη υποστήριξη κανόνα = 10 και ελάχιστη εμπιστοσύνη = 40 παρουσιάστηκαν 7 κανόνες όπως φαίνεται στο σχήμα.



Σχήμα 36. Κανόνες συσχέτισης με βάση τη μέθοδο Apriori.

Γενικευμένη επαγωγή του κανόνα : Η μέθοδος αυτή με ελάχιστη υποστήριξη κανόνα = 10 και με ελάχιστη εμπιστοσύνη κανόνα = 50 οδήγησε σε δύο κανόνες όπως φαίνεται από το παρακάτω σχήμα.



Σχήμα 37. Κανόνες συσχέτισης με βάση τη μέθοδο της γενικευμένης επαγωγής κανόνα.

Η σύγκριση των παραπάνω κανόνων οδηγεί στο συμπέρασμα ότι η ισχύς των 1 , 2 επιβεβαιώνεται εκθέτοντας ακόμη και την αυξανόμενη υποστήριξη. *AN* διαταγή πληρωμής ΙΚΑ *TOTE* διαταγή πληρωμής ΦΠΑ (confidence=81). (Rule 5) – *AN* διαταγή πληρωμής ΦΠΑ *TOTE* διαταγή πληρωμής ΙΚΑ (confidence= 51). (Rule 6)

Συμπεράσματα : Η βασική έκβαση είναι ότι οι διαταγές πληρωμής Φ.Π.Α. και οι διαταγές πληρωμής Ι.Κ.Α. είναι οι δημοφιλέστερες και οι πιο έντονα διασυνδεδεμένες. Η ανίχνευση τέτοιων σχέσεων προσφέρει σε μια τράπεζα λεπτομερή ανάλυση που μπορεί να χρησιμοποιηθεί ως σημείο αναφοράς για την συντήρηση του όγκου των πελατών αρχικά, αλλά και για την προσέγγιση των νέων ομάδων πελατών (επιχειρήσεις) που διευθύνουν αυτές τις πληρωμές προκειμένου να αυξηθεί ο αριθμός των κερδών και των πελατών.

Παρόμοιες καταστάσεις μπορούν να παραχθούν μετά από την ανάλυση των τύπων πληρωμής συγκεκριμένων ομάδων πελατών ως αποτέλεσμα των διαφόρων κριτηρίων, περιοχή κατοίκων, ηλικία, επάγγελμα καθώς και άλλων κριτηρίων που η τράπεζα κρίνει σημαντικά. Οι κανόνες που προκύπτουν παρουσιάζουν σαφώς τις τάσεις που δημιουργούνται στις διάφορες ομάδες πελατών που βοηθούν την τράπεζα για να πλησιάσουν αυτές τις ομάδες καθώς επίσης και για να ξανασχεδιάσουν τις προσφερθείσες ηλεκτρονικές υπηρεσίες, προκειμένου να γίνουν ανταγωνιστικότεροι. Εκτός από τις δύο μεθόδους που χρησιμοποιούνται, ο εντοπισμός των κανόνων των κανόνων συσχέτισης μπορεί να υιοθετήσει άλλες μεθόδους ανάσυρσης δεδομένων όπως τα δέντρα απόφασης προκειμένου να επιβεβαιωθεί η ισχύς των αποτελεσμάτων Σε αυτή την έρευνα η καθιέρωση ενός προτύπου πρόβλεψης σχετικά με τον αριθμό πληρωμών που διευθύνονται μέσω του Διαδικτύου σε σχέση με τον αριθμό των ενεργών χρηστών ερευνάται μαζί με τη δοκιμή της ακρίβειας της

ΕΠΙΛΟΓΟΣ

Πολλές επιχειρήσεις ξέρουν ότι όσο περισσότερη δοσοληψία έχουν με έναν πελάτη τόσο πιο πιστός θα είναι αυτός και ως εκ τούτου πιο επικερδής για την ίδια την επιχείρηση. Έτσι οι τράπεζες και οι ασφαλιστές θέλουν να αυξήσουν τον αριθμό λογαριασμών, τα εμπορικά καταστήματα τον αριθμό προϊόντων που πωλούνται σε μία οικογένεια, επιχειρήσεις πιστωτικών καρτών τον αριθμό συναλλαγών στη κάρτα, οι τηλεφωνικές επιχειρήσεις τον αριθμό πρόσθετων υπηρεσιών κτλ. Η ευκαιρία εσώ είναι το ποια προϊόντα θα πουληθούν μαζί έται ώστε να διερευνηθούν οι σχέσεις μεταξύ πελάτη-επιχειρήσεις. Μόλις καθορίσει η Εξόρυξη Δεδομένων πια προϊόντα θα πρέπει να πωληθούν μαζί, θα πρέπει να είμαστε βέβαιοι ότι αυτά τα προϊόντα θα πωληθούν σε μεγάλες ποσότητες.

Η πληροφορία είναι η νέα δύναμη που οδηγεί τις εξελίξεις στις επιχειρήσεις. Η Εξόρυξη Δεδομένων προσπαθεί να εκμεταλλευτεί τη δύναμη της πληροφορίας και το μετασχηματισμό της σε αποτέλεσμα που μπορούν να φανούν χρήσιμα για την επιχείρηση. Ακριβώς όπως το νερό γύρισε τις ρόδες που έκαναν τις μηχανές σε έναν μύλο να λειτουργήσουν, έτσι και τα δεδομένα πρέπει να συγκεντρωθούν και να διαδοθούν σε όλο τον οργανισμό για να παρέχουν αξιόλογα αποτελέσματα και συμπεράσματα. Εάν η πληροφορία είναι το νερό σε αυτήν την αντιστοιχία, τότε η Εξόρυξη Δεδομένων είναι η ρόδα διαδίδοντας τη δύναμη της πληροφορίας σε όλες τις επιχειρησιακές διαδικασίες.

Στον επιχειρησιακό κόσμο, η Εξόρυξη Δεδομένων παρέχει μία πλήρως καινούργια ικανότητα, τη δυνατότητα να βελτιστοποιηθεί η λήψη αποφάσεων χρησιμοποιώντας τις αυτοματοποιημένες μεθόδους που μαθαίνουν από παρελθοντικές ενέργειες. Τα τελευταία χρόνια, η τεχνολογία συγκλίνει στο να επιτρέψει αυτήν την ικανότητα, και στα επόμενα έτη, πιο σύγχρονα και βελτιστοποιημένα εργαλεία, υλικού και λογισμικού, θα συνεχίσουν αυτή την τάση. Οι επεξεργαστές παράλληλης επεξεργασίας θα γίνουν περισσότερο κοινοί και λιγότερο ακριβοί. Το λογισμικό γίνεται όλο και περισσότερο ικανό για να εκμεταλλευτεί τα παράλληλα και κατανεμημένα συστήματα για τον προσδιορισμό των χρήσιμων προτύπων στα

δεδομένα. Συνεπώς πολλές επιχειρήσεις θα προωθούν ή θα διακινούν τα προϊόντα τους χρησιμοποιώντας τις νέες αυτές τεχνικές.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Εξόρυξη Γνώσης από Βάσεις Δεδομένων. Μ. ΒΑΖΙΡΓΙΑΝΝΗΣ – Μ. ΧΑΛΚΙΔΗ τυπωθήτω-ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ. Έτος έκδοσης 2004. Εκδότης GUTENBERG.

2. Applications of Data Mining –Association Rules. Susan Craw-Nirmalie Wiratunga, Witten & Frank, 2000.

3. J. S. Park, M.-S. Chen, and P. S. Yu, 1995 SIGMOD, pp. 175-186

4. M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. Ullman "Computering Iceberg Queries Effeciently"

5. Ashok Savasere, Edward Omiecinski, and Shamkant Navathe."An efficient algorithm for mining association large databases". In Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95), pages 432-444, Zurich, Switzerland, 1995.

6. H. Toivonen, "Sampling Large Databases for Association Rules", VLDB 1996, pp. 134-145)

7. R. Agrawal, T. Imielinski, A. Swami, "Mining Associations between Sets of Items in Massive Databases", ACM SIGMOD Int'l Conference on Management of Data, Washington D.C., May 1993, 207-216

8. R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. Of the 20th Int'l Conference of Databases, Santiago, Chile, Sept. 1994

9. Hand D, Mannila H and Smyth P (2001). Principles of Data Mining. MIT press, Massachusetts

10. Inroduction to data mining and knowledge discovery. 3rd edition. Two Crows Corporation (http://www.twocrows.com)