

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
61740
005.741
ΦΕΛ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**«Τεχνικές Οπτικής Παρουσίασης των Αποτελεσμάτων των
Διαδικασιών Εξόρυξης Γνώσης»**

**Φελούκας Παναγιώτης
Μ3970008**

Επιβλέπων Καθηγητής: Μιχάλης Βαζιργιάννης

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ**



000000 377744



Ευρετήριο

Executive Summary

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
εισ.
Αρ. 6140
ταξ. 005.741
φελ

Κεφάλαιο 1: Εξόρυξη Γνώσης Από Μεγάλες Βάσεις Δεδομένων (Data Mining) 1

1.1 Εισαγωγή	1
1.2 Ορισμός.....	3
1.3 Η Μεθοδολογία Εξόρυξης Γνώσης.....	4
1.4 Αλγόριθμοι Εξόρυξης Γνώσης.....	6
1.5 Τεχνικές Πρόβλεψης vs Τεχνικών Επιβεβαίωσης.....	7
1.6 Προβλήματα & Δυσκολίες Ανάπτυξης Συστημάτων Εξόρυξης Γνώσης	9
1.7 Data Mining & Data Warehouse.....	10
1.7.1 Ορισμός του Data Warehouse	10

Κεφάλαιο 2: Τεχνικές Οπτικής Παρουσίασης των Μοντέλων Data Mining 12

2.1 Γιατί Χρησιμοποιούμε Τεχνικές Οπτικής Παρουσίασης.....	12
2.2 Αρχές και Κανόνες που Διέπουν τις Τεχνικές Οπτικής Παρουσίασης	14
2.3 Τεχνικές Οπτικής Παρουσίασης	16
2.3.1 Τεχνικές Μείωσης Διαστάσεων	17
MDS (Multi Dimensional Scaling).....	18
Αλγόριθμος Karhunen-Loeve.....	18
Ανάκτηση και Κατηγοριοποίηση (Retrieval και Clustering)	18
Αλγόριθμος Fast-Map.....	19
2.3.2 Γεωμετρικές Τεχνικές (Geometric)	23
Τεχνική Προβολών (Projection Views)	23
Παράλληλη Γραμμική Αναπαράσταση (Parallel coordinate)	23
Πίνακες από Scatterplots	24
Τοπογράφημα (Landscapes).....	26
Δισδιάστατα Δείγματα του κ Διάστατου Χώρου (Hyperslice)	26
2.3.3 Τεχνικές Βασιζόμενες σε Εικόνες	27
Πρόσωπα Chernoff	27
Γραμμικά Σχήματα (Stick Figures).....	28
Μορφοποίηση Σχήματος (Shape Coding)	29
Εγχρωμη Μορφοποίηση Σχήματος (Color Icons)	30
Τεχνική Αναζήτησης Κειμένων – TileBars (Πλακάκια)	30
2.3.4 Ιεραρχικές Τεχνικές	32
Ιεραρχική Κατανομή (Dimensional Stacking)	32
Ενθυλακωμένα Διαγράμματα – Worlds within Worlds	33
Δεντρική - Ενθυλακωμένη Παράθεση (Treemap – Venn)	35
Κωνικά Δέντρα (Cone Trees).....	38
Ιεραρχικός Κύβος (InfoCube)	39
Τεχνική Εξερεύνησης Δομής Καταλόγων - FSN (File System Navigation)	40
2.3.5 Τεχνικές Σχεδίασης σε Επίπεδο Pixels	41
Τεχνικές Ανεξάρτητες Ερωτήσεων (Query Independent)	42
Τεχνικές Εξαρτώμενες από Ερωτήσεις (Query Dependent)	45
Τεχνικές Ομαδοποίησης (Grouping)	47



2.3.6 Τεχνικές Γραφημάτων (Graph Based)	48
Ορθογώνιο Γράφημα	49
Συμμετρικό Γράφημα	49
Γράφημα Κατηγοριοποίησης (Cluster Based).....	49
Κατευθυνόμενο - ακυκλικό γράφημα (Acylic Graph)	49
Γραφήματα Υπέρ-Δεντρικών Δομών (Hygraphs)	51
Γραφήματα SeeNet	52
Γράφημα Narcissus.....	53
2.3.7 Τεχνικές Παραμόρφωσης Εικόνας (Distortion).....	54
Τοίχος με Προοπτική (Perspective Wall).....	54
Τεχνική Φακού Μεγέθυνσης σε Πίνακες (Table Lens)	56
Τεχνική με Προοπτική Φακού Μεγέθυνσης (Fisheye View)	57
Τεχνική Δέντρων «Υπερβολικής» Γεωμετρίας (Hyperbolic Trees)	58
Τεχνική Υπερ-Κύβου (HyperBox).....	61
2.3.8 Δυναμικές Τεχνικές	61
2.4 Αξιολόγηση και Σύγκριση.....	62
 Κεφάλαιο 3: Σύστημα Εξόρυξης Γνώσης.....	65
3.1 Εισαγωγή	65
3.2 Περιγραφή Συστήματος	66
3.3 Προτεινόμενη Λύση	69
3.3.1 Clustering.....	69
1 Διάσταση.....	70
2 ή 3 Διαστάσεις.....	71
Κ Διαστάσεις.....	72
3.3.2 Classification.....	73
3.3.3 Association Rules	75
 Πίνακας Εικόνων.....	i
Αναφορές	iv



EXECUTIVE SUMMARY

Η μεγάλη ανάπτυξη των εφαρμογών πληροφορικής στον επιχειρηματικό και ερευνητικό τομέα οδήγησε στην δημιουργία πολύ μεγάλων βάσεων όπου φυλάσσονταν διαφόρων ειδών στοιχεία. Ωστόσο, η πληροφορία που μπορούσε να εξαχθεί από αυτά ήταν πολύ φτωχή π.χ. αναζήτηση εγγραφών που πληρούν κάποιες συνθήκες. Έτσι, δημιουργήθηκε η ανάγκη καλύτερης εκμετάλλευσης των διαφόρων βάσεων, με τρόπο ώστε να μπορεί να εξαχθεί γνώση. Η αρχιτεκτονική που δημιουργήθηκε (Data Mining – Εξόρυξη Γνώσης) περιλαμβάνει τα ακόλουθα βήματα: προσδιορισμός των στόχων, δημιουργία ενός συνόλου δεδομένων (το οποίο πιθανώς να ενοποιεί διάφορες πρωτογενής πηγές δεδομένων), καθαρισμός και προεπεξεργασία των δεδομένων, περιορισμός του όγκου τους, εφαρμογή αλγορίθμων εξόρυξης γνώσης, οπτική παρουσίαση των αποτελεσμάτων, αξιολόγηση των αποτελεσμάτων.

Το κομμάτι της παραπάνω αρχιτεκτονικής με το οποίο ασχολείται η παρούσα εργασία είναι το προτελευταίο, δηλ. η οπτική παρουσίαση των αποτελεσμάτων. Είναι ένα από τα κρισιμότερα μέρη της διαδικασίας γιατί αναλαμβάνει την αλληλεπίδραση με το χρήστη και γιατί σ' αυτό το τμήμα παρουσιάζεται όλη η δουλειά που έχει γίνει από τα προηγούμενα βήματα. Συνεπώς, αν η οπτική παρουσίαση δεν είναι **κατανοητή και αξιοπιστή** τότε ανεξάρτητα από την ποιότητα και την αποτελεσματικότητα των αλγορίθμων που έχουν εφαρμοστεί για την εξαγωγή της γνώσης, το σύστημα έχει μεγάλες πιθανότητες αποτυχίας.

Τρεις παράγοντες προσδιορίζουν τον βαθμό κατανόησης των αποτελεσμάτων:

1. Παρουσίαση: πρέπει να δοθεί μεγάλη βαρύτητα στην **ισοστάθμιση** των παραγόντων: **πολυπλοκότητα/απώλεια στοιχείων**. Το γράφημα θα πρέπει να είναι απλό ώστε να συμβαδίζει με τον τρόπο σκέψης του χρήστη καθώς και να μην χάνει σε ποιότητα.
2. Αλληλεπίδραση: Ο χρήστης πρέπει να έχει τη δυνατότητα επικοινωνίας με το οπτικό αποτέλεσμα. Θα πρέπει να λειτουργεί μια διαδικασία επανατροφοδότησης στοιχείων (από το χρήστη προς το σύστημα) ώστε μια νέα ερώτηση να είναι επέκταση της παλιάς, και μάλιστα αυτό να γίνεται άμεσα (π.χ. Drill Down, Roll Up, Slice & Dice κ.λ.π.)
3. Ολοκλήρωση: Να μπορεί το οπτικό αποτέλεσμα να δώσει μια **ολοκληρωμένη εικόνα** στον χρήστη για τα δεδομένα του.

Όσον αφορά την αξιοπιστία, αυτή αποκτιέται μετά από μακροχρόνια χρήση του συστήματος με δεδομένο ότι δεν οδηγεί σε λάθος συμπεράσματα. Ωστόσο υπάρχουν διάφοροι άλλοι παράγοντες, οι οποίοι μπορούν να πείσουν για την αξιοπιστία του μοντέλου.

1. Είναι απαραίτητο το μοντέλο να δηλώνει με σαφήνεια μέσω της οπτικής παρουσίασης το βαθμό εμπιστοσύνης των αποτελεσμάτων. Κάθε στατιστική διαδικασία ανεύρεσης συσχετίσεων ή τάσεων μέσα στα δεδομένα εξάγει ένα



σύνολο αποτελεσμάτων τα οποία χαρακτηρίζονται από ένα ποσοστό εμπιστοσύνης.

2. Το κάθε πρόγραμμα έχει κάποιους περιορισμούς, τους οποίους δεν πρέπει να τους κρύβει από τον χρήστη, αλλά να τους καθορίζει διακριτικά, ώστε ο χρήστης να μην ξεπερνάει τα όρια χρήσης του, οδηγώντάς το σε ανακριβή αποτελέσματα.
3. Η παρουσίαση των αποτελεσμάτων θα πρέπει να είναι σαφής και κατανοητή αποφεύγοντας σύνθετα αποτελέσματα. Πολλές φορές ένα δισδιάστατο ραβδόγραμμα αποδεικνύεται καλύτερο από ένα πολυδιάστατο γράφημα.
4. Το αποτέλεσμα πρέπει να πλησιάζει τον τρόπο σκέψης του χρήστη.

Η αναζήτηση υλοποιημένων αλλά και ερευνητικών τεχνικών οδηγησε στην δημιουργία μίας μεγάλης λίστας η οποία δηλώνει σαφώς την μεγάλη ανάγκη αλλά και την προσπάθεια που καταβάλλεται στο χώρο αυτό για ανεύρεση τεχνικών που θα διευκολύνουν την εξερεύνηση και κατανόηση των αποτελεσμάτων των διαδικασιών εξόρυξης γνώσης. Πολλές από τις τεχνικές αυτές παρουσιάζουν κοινά χαρακτηριστικά ή απευθύνονται στον ίδιο τύπο δεδομένων, γι' αυτό δημιουργήθηκαν οι ακόλουθες κατηγορίες:

1. Κατηγοριοποίηση των δεδομένων (Classification)
2. Οπτική παραμόρφωση του αποτελέσματος (Distortion)
3. Δυναμικές τεχνικές (Dynamic).

Οι τεχνικές της πρώτης κατηγορίας είναι στην πράξη και πιο ενδιαφέρουσες και πρωτότυπες. Στην κατηγορία αυτή αναλύονται τεχνικές, όπως προβολές τημάτων του τρισδιάστατου χώρου στον δισδιάστατο, συστήματα παράλληλων συντεταγμένων, πίνακες από scatter plots για παρουσίαση πολυδιάστατων δεδομένων, k-διάστατοι πίνακες από φέτες δεδομένων από τον πολυδιάστατο χώρο, τεχνικές που βασίζονται σε χρήση εικόνων ή συμβόλων για αναπαράσταση της πολυδιάστατης πληροφορίας, προσανατολισμένες σε pixel τεχνικές για παρουσίαση μεγάλων όγκων δεδομένων, τεχνικές που χρησιμοποιούνται για αναζήτηση κειμένων, τεχνικές παρουσίασης δεδομένων που έχουν iεραρχική δομή και τρισδιάστατα και δισδιάστατα γραφήματα.

Ο μεγάλος αριθμός των τεχνικών αυτής της κατηγορίας οδηγεί σε ανάγκη περαιτέρω κατηγοριοποίησής της:

1. Κατηγοριοποίηση των δεδομένων (Classification)
 - i. Μείωση διαστάσεων: Σκοπός των τεχνικών αυτών είναι η παρουσίαση k-διάστατων δεδομένων σε d-διάστατο χώρο, όπου $d \ll k$ για εύκολη αναπαράστασή τους. Μια πρόσφατη προσπάθεια σε τέτοιου είδους τεχνικές είναι ο αλγόριθμος FastMap [23]



- ii. Γεωμετρικές: Οι γεωμετρικές τεχνικές αποσκοπούν στην εύρεση προβολών πολυδιάστατων δεδομένων σε δισδιάστατο χώρο και στην επιλογή εκείνων που δείχνουν ενδιαφέρουσες. Τέτοιες τεχνικές είναι οι:
- τεχνική προβολών (projection views),
 - παράλληλη γραμμική αναπαράσταση (parallel coordinate),
 - πίνακες από scatterplots,
 - τοπογράφημα (landscapes),
 - δισδιάστατα δείγματα του k διάστατου χώρου (hyperslice)
- iii. Βασιζόμενες σε εικόνες: Η βασική ιδέα αυτών των τεχνικών είναι η παρουσίαση των αποτελεσμάτων χρησιμοποιώντας εικονίδια. Κάθε μονάδα των δεδομένων αντιπροσωπεύεται από κάποιο εικονίδιο, το οποίο έχει επιλεγεί με βάση κάποιους συγκεκριμένους κανόνες ώστε να εκφράζει τα χαρακτηριστικά της κάθε μονάδας. Υπάρχουν διάφορες παραλλαγές των τεχνικών αυτών:
- Πρόσωπα Chernoff,
 - Γραμμικά Σχήματα (Stick Figures),
 - Μορφοποίηση Σχήματος (Shape Coding),
 - Έγχρωμη Μορφοποίηση Σχήματος (Color Icons),
 - Τεχνική Αναζήτησης Κειμένων – TileBars (Πλακάκια)
- iv. Βασιζόμενες σε pixels: Η τεχνική αυτή συνίσταται στην παρουσίαση κάθε τιμής ενός γνωρίσματος ενός πίνακα με ένα pixel στην οθόνη χρησιμοποιώντας ένα συγκεκριμένο χρωματισμό, ο οποίος παρουσιάζει τη σχετικότητα (relevance) των δεδομένων. Κάθε γνώρισμα του πίνακα παρουσιάζεται σε διαφορετικά υπό-παράθυρα στην οθόνη. Οι τεχνικές αυτές κατηγοριοποιούνται σε 3 είδη:

- Query dependent που παρουσιάζουν τα δεδομένα βασιζόμενες σε ερωτήσεις του χρήστη.

Οι αλγόριθμοι, οι οποίοι χρησιμοποιούνται χωρίζονται σε

1. Snake Spiral
2. Snake Axes

- Query independent που απλώς παρουσιάζουν τα δεδομένα. Για να λειτουργήσουν τέτοιου είδους τεχνικές θα πρέπει να υπάρχει μέσα στα δεδομένα μια λογική ταξινόμησης π.χ. time series. Αλγόριθμοι για αυτή την κατηγορία είναι οι εξής:

1. Screen Filling Curve



2. Recursive pattern

- Grouping techniques: Βασική ιδέα είναι η συνένωση των πολλών παραθύρων που χρησιμοποιούνται για κάθε μεταβλητή σε ένα παράθυρο, ομαδοποιώντας τα δεδομένα για κάθε data item, δηλ για κάθε ξεχωριστή τιμή κάθε στήλης. (Εικόνα 23).
- v. Ιεραρχικές: Οι τεχνικές αυτές παρουσιάζουν τα δεδομένα, τα οποία εμπεριέχουν κάποιου είδους ιεραρχική δομή, χρησιμοποιώντας μία ιεραρχική κατηγοριοποίηση της οθόνης σε υπό-τμήματα. Για παράδειγμα, μία βάση δεδομένων, η οποία περιέχει πληροφορίες για τον πληθυσμό της Ελλάδας σε επίπεδο Πόλης, Επαρχίας, Νομού και Διαμερίσματος είναι ιεραρχικά δομημένη μιας και όλα τα επίπεδα που ανέφερα προηγουμένως συνθέτουν μία διαδρομή εξερεύνησης στοιχείων που αφορούν την Ελλάδα εμβαθύνοντας σταδιακά. Τέτοιες τεχνικές είναι οι ακόλουθες:
 - Ιεραρχική Κατανομή (Dimensional Stacking)
 - Ενθυλακωμένα Διαγράμματα – Worlds within Worlds
 - Δεντρική - Ενθυλακωμένη Παράθεση (Treemap – Venn)
 - Κωνικά Δέντρα (Cone Trees)
 - Ιεραρχικός Κύβος (InfoCube)
 - *Τεχνική Εξερεύνησης Δομής Καταλόγων - FSN (File System Navigation)*
- vi. Γραφήματα: Μία τελευταία κατηγορία γεωμετρικών τεχνικών είναι τα γραφήματα. Η έννοια γράφημα χρησιμοποιήθηκε πολλές φορές στις μέχρι τώρα αναφερθείσες τεχνικές αλλά δεν πρέπει να συγχέεται με τις τεχνικές γραφημάτων. Οι τεχνικές, οι οποίες βασίζονται σε γραφήματα έχουν ορισμένα χαρακτηριστικά που τις διαχωρίζουν από τις υπόλοιπες. Τέτοια χαρακτηριστικά είναι:
 - Δεν παρουσιάζουν πληροφορία σε σύνολα αξόνων. Έχουν τη δυνατότητα να αναπαραστήσουν πολυδιάστατη πληροφορία αλλά χωρίς τη χρήση τρισδιάστατων αξόνων και αυτό τις διαχωρίζει από τις γεωμετρικές τεχνικές με τις οποίες θα μπορούσε κάποιος να τις συσχετίσει.
 - Αποτελούνται από σύνολα γραμμών-ακμών (ευθέων ή καμπύλων) τα οποία ενώνουν σημεία-κόμβους, σχηματίζοντας έτσι πολυγωνικούς σχηματισμούς γραμμών.
 - Απευθύνονται κυρίως σε δεδομένα που εμπεριέχουν πληροφορίες συσχέτισης και αλληλεξάρτησης (relational information).
 - Υπάρχουν γραφήματα 2 και τριών διαστάσεων. Η εισαγωγή τρισδιάστατης τεχνολογίας πραγματώνεται σε ήδη υπάρχουσες δισδιάστατες τεχνικές με σκοπό την καλύτερη εκμετάλλευση



χώρου, χωρίς αυτό να αλλοιώνει την υφή του αποτελέσματος (λογική και σημασιολογική).

- Τα γραφήματα δίνουν μεγάλη έμφαση σε ένα παράγοντα, οποίος δεν επισημάνθηκε ιδιαίτερα από τις άλλες τεχνικές. Ο παράγοντας αυτός ονομάζεται "Αισθητική" και σημαίνει τις λεπτές αλλά σημαντικές αρχές που καθορίζουν τη σχεδίαση ενός γραφήματος. Η αλλαγή αυτών των κανόνων αισθητικής οδηγεί σε διαφορετικές προσεγγίσεις και κατά συνέπεια δημιουργεί νέους τύπους γραφημάτων.

Οι τεχνικές οπτικής παραμόρφωσης του αποτελέσματος, εισάγουν μία νέα λογική στην παρουσίαση δεδομένων παραμορφώνοντας το αποτέλεσμα. Όπως, αναφέρθηκε, οι διαδικασίες εξόρυξης γνώσης εφαρμόζονται σε μεγάλες βάσεις δεδομένων, συνεπώς είναι πολύ πιθανό τα αποτελέσματα που πρέπει να παρουσιαστούν να είναι πολλά. Το κυριότερο πρόβλημα που αντιμετωπίζουν οι τεχνικές που εφαρμόζονται σε τέτοια δεδομένα είναι αδυναμία παρουσίασης των δεδομένων στο σύνολό τους, χωρίς να δημιουργούν ακατανόητα αποτελέσματα. Εφαρμόζοντας τεχνικές παραμόρφωσης της εικόνας βελτιστοποιείται το οπτικό αποτέλεσμα παρουσιάζοντας ένα μεγάλο μέρος του συνόλου των δεδομένων, εστιάζοντας σε αυτά που αναμένεται να έχουν μεγαλύτερη σημασία. Οι τεχνικές αυτές χρησιμοποιούν τεχνικές από την πρώτη κατηγορία και συνδυασμούς τους εφαρμόζοντας παράλληλα τη δική τους οπτική γωνία παρουσίασης. Τέλος, στην κατηγορία αυτή ανήκουν και οι τεχνικές:

- Τοίχος με Προοπτική (Perspective Wall)
- Τεχνική Φακού Μεγέθυνσης σε Πίνακες (Table Lens)
- Γράφημα με Προοπτική Φακού Μεγέθυνσης (Fishey View)
- Τεχνική Δέντρων «Υπερβολικής» Γεωμετρίας (Hyperbolic Trees)

Οι Δυναμικές τεχνικές αφορούν υλοποιημένες τεχνικές οπτικής παρουσίασης, οι οποίες ανήκουν στις άλλες δύο κατηγορίες και για τις οποίες έχει δημιουργηθεί ένα δυναμικό περιβάλλον αλληλεπίδρασης με το χρήστη.

Τελικός, σκοπός της εργασίας αυτής είναι να επιλεγούν οι τεχνικές εκείνες που είναι καταληλότερες για την παρουσίαση των αποτελεσμάτων ενός συστήματος εξόρυξης γνώσης, το οποίο υλοποιεί αλγόριθμους classification, clustering και association. Έτσι, για τον σχεδιασμό του τμήματος οπτικοποίησης του συστήματος χρησιμοποιήθηκε η λογική των pixel τεχνικών, η τεχνολογία των γραφημάτων και η τεχνολογία των διαγραμμάτων σε 2 και 3 διαστάσεις. Συγκεκριμένα:



➤ Clustering

➤ 1 Διάσταση:

Οι Pixel oriented τεχνικές ενδείκνυνται για παρουσίαση κατηγοριών στα δεδομένα μίας διάστασης. Χάρη στους αλγόριθμους γεμίσματος της οθόνης, Peano Hilbert & Morton, που χρησιμοποιούνται δημιουργείται μία αρκετά καλή και σαφή εικόνα του αποτελέσματος με ευκρινή ένδειξη των κατηγοριών, ειδικά όταν χρησιμοποιούνται χρώματα. Πιθανή ύπαρξη ιεραρχικής πληροφόρησης θα μπορούσε να αναπαρασταθεί με χρήση των τεχνικών recursive pattern. Τέλος, το χρώμα θα έδειχνε το βαθμό συμμετοχής του κάθε αντικειμένου στην κάθε κατηγορία.

➤ 2 – 3 Διαστάσεις:

Για αναπαράσταση της πληροφορίας σε αυτό το επίπεδο προτείνεται η χρήση των τεχνικών Hyperslices, Dimensional Stacking και τρισδιάστατων διαγραμμάτων.

➤ K – Διαστάσεις

Τέλος, για την περίπτωση όπου δεν υπάρχει ιεραρχική δομή στα δεδομένα και θέλουμε να αναπαραστήσουμε περισσότερες των τριών διαστάσεων τότε προτείνεται η χρήση pixel oriented τεχνικών. Συγκεκριμένα, για κάθε διάσταση-γνώρισμα του πίνακα που αναπαρίσταται θα χρησιμοποιείται ένα παράθυρο της οθόνης. Η κάθε κατηγορία θα αντιστοιχίζεται με ένα χρώμα, το οποίο θα παρουσιάζεται στον χρήστη με χρήση μίας μπάρας σε κάποιο σημείο της οθόνης. Εκεί, θα αναγράφονται και οι τιμές ή το εύρος των τιμών για τις οποίες σχηματίζεται η κάθε κατηγορία. Η τεχνική γεμίσματος των παραθύρων θα είναι η τεχνική γεμίσματος από αριστερά προς τα δεξιά και ανάστροφα. Μεγάλο πλεονέκτημα αυτής της τεχνικής είναι ότι μπορούν να παρουσιαστούν και κατηγορίες στις οποίες συμμετέχουν λιγότερα των k γνωρισμάτων του πίνακα (που σχηματίζουν τις διαστάσεις).

➤ Classification

Μία τεχνική που μπορεί να χρησιμοποιηθεί σ' αυτήν την περίπτωση είναι η Treemap. Η τεχνική αυτή χρησιμοποιείται κυρίως για ιεραρχικά δεδομένα, αλλά αν θεωρήσουμε ότι υπάρχουν 2 επίπεδα ιεραρχίας: Γνώρισμα → Κατηγορία, τότε η χρήση της είναι δυνατή και οδηγεί σε γρήγορες αποφάσεις και συμπεράσματα. Μία οπτική βελτίωση της μεθόδου είναι η χρήση χρωμάτων για την αναπαράσταση των κατηγοριών καθώς και τρισδιάστατα εφέ.

➤ Association

Για την οπτική παρουσίαση των κανόνων συσχέτισης καλύτερη τεχνική είναι τα γραφήματα. Το σύστημα που έχει υλοποιηθεί παρέχει απλούς κανόνες 1-1, δηλ. Στοιχείο A → Στοιχείο B, το οποίο σημαίνει ότι το στοιχείο A παρουσιάζει κάποια



συσχέτιση με το στοιχείο B, με κάποιο confidence και κάποιο support. Η μετρική Support δείχνει το ποσοστό εμφάνισης επί το σύνολο των δεδομένων του στοιχείου A, ενώ η μετρική confidence δείχνει το ποσοστό των εγγραφών που παρουσιάζουν τη συσχέτιση A → B από το σύνολο των εγγραφών όπου υπάρχει το Στοιχείο A. Οι μετρικές αυτές παρέχονται για κάθε κανόνα συσχέτισης και στην ουσία παρουσιάζουν τον βαθμό εμπιστοσύνης



ΚΕΦΑΛΑΙΟ Ι^ο

**Εξόρυξη Γνώσης Από Μεγάλες Βάσεις
Δεδομένων (Data Mining)**



Κεφάλαιο1: Εξόρυξη Γνώσης Από Μεγάλες Βάσεις Δεδομένων (Data Mining)

1.1 Εισαγωγή

Οι πρώτες βάσεις δεδομένων σχεδιάστηκαν έτσι ώστε να μπορούν να αποθηκεύουν και ανακτούν δεδομένα. Η διαδικασία αυτή ήταν σχετικά απλή και δεν επιζητούσε ιδιαίτερες τεχνολογίες για την υλοποίησή της. Με την αύξηση του όγκου των αποθηκευόμενων δεδομένων παρουσιάστηκε η ανάγκη (κυρίως από υψηλά στελέχη επιχειρήσεων) επεξεργασίας των δεδομένων με σκοπό να πάρουν κάποιο είδος πληροφορίας. Ετσι, εμφανίστηκαν οι πρώτες προσπάθειες δημιουργίας εργαλείων που θα εξυπηρετούσαν αυτό το σκοπό. Τέτοιο εργαλείο είναι η γλώσσα SQL.

Με την πάροδο του χρόνου και την τεχνολογική βελτίωση του υλικού (Hardware) και του λογισμικού (Software) εμφανίστηκαν τα πρώτα συστήματα OLAP. Τα συστήματα αυτά απαιτούσαν ιδιαίτερη οργάνωση της βάσης δεδομένων προκειμένου να μπορέσουν να λειτουργήσουν. Συγκεκριμένα, για την υποστήριξη αυτών των εργαλείων δημιουργήθηκαν διάφορες μεθοδολογίες και πρότυπα οργάνωσης των βάσεων δεδομένων. Οι νέες δομές στις βάσεις δεδομένων ονομάστηκαν Data WareHouse και σκοπός τους ήταν όχι η αποθήκευση δεδομένων αλλά η οργάνωσή τους σε συγκεκριμένα σχήματα και η ολοκλήρωση (integration) όλων των υπαρχόντων βάσεων δεδομένων σε μια δομή, η οποία θα ήταν εύκολα επεξεργάσιμη.

Σκοπός της τεχνολογίας αυτής ήταν η συγκριτική και συγκεντρωτική παρακολούθηση των δεδομένων προκειμένου να βγάλουμε "γενικά" συμπεράσματα για την επιχείρηση, τον οργανισμό κ.λ.π. Το μεινέκτημα της τεχνολογίας αυτής είναι ότι ο σχεδιαστής του συστήματος πρέπει να ερευνήσει τη βάση και να προσδιορίσει ποια θα είναι τα κρίσιμα σημεία στα οποία πρέπει να δοθεί ιδιαίτερη σημασία και για τα οποία θα πρέπει να παρέχεται πληροφορία χρήσιμη στην επιχείρηση.

Το μεινέκτημα αυτό λύνει η τεχνολογία '**Data Mining**' κάνοντας χρήση μεθόδων από τον κλάδο της τεχνητής νοημοσύνης και της στατιστικής.

Ακολουθεί ένας πίνακας (πίνακας 1) ο οποίος παρουσιάζει την εξελικτική διαδικασία στη χρήση όλο και πιο έξυπνων τεχνολογιών για την επεξεργασία των βάσεων δεδομένων μέχρι σήμερα:



{PRIVATE} Εξέλιξη	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"Ποιος είναι ο συνολικός μισθός μου τα τελευταία 5 χρόνια?"	Computers, δισκέτες, δίσκοι	IBM, CDC	Παροχή στατικής πληροφορόρησης και στατικής επεξεργασίας
Data Access (1980s)	"Που κυμαίνονται οι πωλήσεις μου τον τελευταίο μήνα?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Δυναμική επεξεργασία των δεδομένων, δυνατότητα χρήσης φυσικής γλώσσας για ερωτήσεις σε βάσεις δεδομένων (SQL)
Data Warehousing & Decision Support (1990s)	Ποιες είναι οι πωλήσεις μου στην Ελλάδα το έτος 1998? Drill Down σε επίπεδο νομού, πόλης και σε επίπεδο μήνα και ημέρας.	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrate gy, SAS Institute	Δυναμική επεξεργασία των δεδομένων με χρήση πολυδιάστατων τεχνικών αναπαράστασης των δεδομένων και on-line επεξεργασία.
Data Mining (Emerging Today)	Θέλω πρόβλεψη για την συμπεριφορά των πελατών αν ακολουθήσω τη στρατηγική 'XXX'	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups, SAS Institute	Proactive πληροφορία, δυνατότητα εξόρυξης γνώσης για στήριξη αποφάσεων.

Πίνακας 1. Βήματα εξέλιξης μέχρι το Data Mining

1.2 Ορισμός

Η τεχνολογία Data Mining είναι μια σχετικά νέο-εφαρμοζόμενη τεχνολογία και αυτός είναι ένας λόγος της μη ύπαρξης σαφούς ορισμού. Στις παρακάτω παραγράφους θα δοθεί μια περιγραφή του όρου 'Data Mining' με σκοπό να γίνει κατανοητή η έννοια του όρου και οι στόχοι που εξυπηρετεί.

Σε πολλές πλέον επιχειρήσεις υπάρχουν μεγάλες βάσεις δεδομένων ψηφιακής πληροφορίας, από τις οποίες έχει υπολογιστεί ότι χρησιμοποιείται μόνο το 7% των δεδομένων τους. Για παράδειγμα, οι πωλήσεις μιας αλυσίδας καταστημάτων, οι κινήσεις των πιστωτικών καρτών, το αρχείο ενός ιατρικού κέντρου, τα αρχεία των τηλεπικοινωνιακών οργανισμών. Η σημερινή τεχνολογία (δυνατοί προσωπικοί υπολογιστές, πολύ καλά προγράμματα για διαχείριση βάσεων δεδομένων) ευνοεί την οικονομική, αποτελεσματική και γρήγορη αποθήκευση και ανάκτηση των δεδομένων. Ωστόσο, σε επίπεδο επιχειρήσεων, η πληροφορία αυτή (επιπέδου εγγραφής δεδομένων) δεν έχει μεγάλη σημασία. Αυτό που θα ήταν μεγάλης σημασίας είναι η γνώση που συνεπάγεται η επεξεργασία των 'row data'. Για παράδειγμα, μια βάση δεδομένων ενός πολυκαταστήματος που έχει καταχωρημένες τις πωλήσεις των προϊόντων ανά πελάτη, θα μπορούσε να μας δώσει χρήσιμη πληροφορία και συσχετίσεις μεταξύ προϊόντων και δημογραφικών ομάδων των πελατών μας. Δηλ. Θα μπορούσε να μας πει ότι οι κάτοικοι της Κηφισιάς που έχουν εισόδημα μεγαλύτερο από X δρχ. ετησίως αγοράζουν την Υ μάρκα αναψυκτικών. Αυτή η πληροφορία είναι άγνωστη αλλά ταυτόχρονα και πολύ χρήσιμη για το τμήμα πωλήσεων που προσδιορίζει τις τακτικές προώθησης των αναψυκτικών.

Ένας ορισμός που θα μπορούσε να δοθεί για την τεχνολογία του Data Mining είναι ο ακόλουθος:

Data Mining είναι η διαδικασία εκείνη που σκοπό έχει την ανακάλυψη άγνωστων (μη τετριμμένων) και ικανών συσχετίσεων, τάσεων και μοντέλων-υποδειγμάτων που κρύβονται σε μεγάλα σύνολα δεδομένων, χρησιμοποιώντας μεθοδολογίες από τους τομείς των μαθηματικών, της στατιστικής και της τεχνητής νοημοσύνης.

1.3 Η Μεθοδολογία Εξόρυξης Γνώσης

Ο όρος Data Mining, όπως αναφέρθηκε παραπάνω, σημαίνει εξόρυξη πολύτιμης πληροφορίας από ένα μεγάλο σύνολο δεδομένων. Ωστόσο, μπορεί να συναντήσουμε τον συναφή όρο 'Knowledge Discovery in Databases' που έχει πιο ολοκληρωμένη έννοια, δίνοντας έμφαση στη συνολική διαδικασία που ακολουθείται για την εξόρυξη των δεδομένων.

Συγκεκριμένα, το Data Mining θεωρείται ένα μέρος της διαδικασίας εξόρυξης γνώσης, δηλ. της διαδικασίας 'Knowledge Discovery in Databases' (KDD). Η διαδικασία KDD προϋποθέτει μερικά βήματα δημιουργίας κατάλληλων δεδομένων τα οποία θα επεξεργαστούν οι μέθοδοι-αλγόριθμοι Data Mining. Τα βήματα αυτά αναλύονται παρακάτω.



Συνήθως, οι όροι Data Mining και KDD χρησιμοποιούνται ως συνώνυμοι !

Η περιγραφή της διαδικασίας εξόρυξης δεδομένων που ακολουθεί δεν είναι αυστηρή ούτε απόλυτη, δηλ. είναι μια ευέλικτη μεθοδολογία η οποία μπορεί να ακολουθηθεί είτε επακριβώς είτε να προσαρμοστεί στις ανάγκες της δικής μας επιχείρησης - οργανισμού.

Η μεθοδολογία αυτή αποτελείται από τα ακόλουθα βήματα:

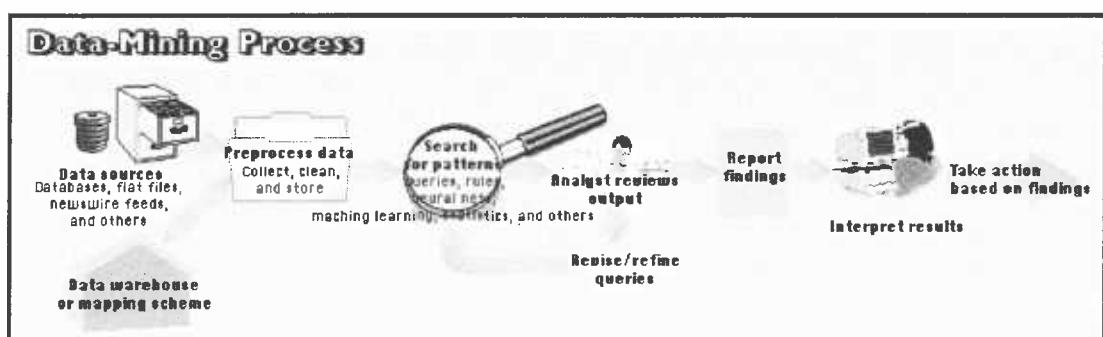
- 1. Προσδιορισμός των στόχων (goals) της εφαρμογής**
- 2. Δημιουργία ενός συνόλου των δεδομένων (data set) το οποίο θα χρησιμοποιηθεί, π.χ. ένα σύνολο από πίνακες οι οποίοι έχουν συγκεντρωμένες τις πληροφορίες πωλήσεων, όλων των υπαλλήλων ανά γεωγραφική περιοχή αλλά για τρία μόνο προϊόντα. Για παράδειγμα η δημιουργία του συνόλου των πινάκων θα μπορούσε να γίνει με χρήση SQL ερωτήσεων στη βάση δεδομένων, επιλέγοντας στήλες από διαφορετικούς πίνακες της βάσης και μερικές ή όλες από τις εγγραφές (subset).**
- 3. Data cleaning και preprocessing.** Περιλαμβάνει όλες τις διαδικασίες με τις οποίες αφαιρούνται τα δεδομένα εκείνα τα οποία δεν προσδίδουν χρήσιμη πληροφορία π.χ. missing values στα πεδία.
- 4. Data reduction.** το στάδιο αυτό έχει σαν σκοπό την μείωση των δεδομένων όσο το δυνατόν περισσότερο δημιουργώντας ομαδοποιήσεις και αφαιρώντας στήλες οι οποίες δεν χρειάζονται. Επειδή, τα δεδομένα τα οποία επεξεργαζόμαστε για την εξαγωγή γνώσης είναι υπερβολικά πολλά προσπαθούμε να τα μειώσουμε (με τις ομαδοποιήσεις, χρήση αθροιστικών αρχείων) προκειμένου να γίνει η διαδικασία ταχύτερη και για να μας δώσει πιο σωστά αποτελέσματα.

5. **Εφαρμογή αλγορίθμων Data Mining.** Στο στάδιο αυτό επιλέγουμε ποιον αλγόριθμο Data Mining θα ακολουθήσουμε πάντα σε σχέση με τους στόχους που έχουμε θέσει στο πρώτο βήμα, και τον εφαρμόζουμε στα δεδομένα που έχουμε δημιουργήσει μετά τις φάσεις 2,3 & 4.
6. **Δημιουργία γραφημάτων.** Αφού εκτελεστούν οι αλγόριθμοι Data Mining επιλέγουμε ένα ή περισσότερους τρόπους παρουσίασης των αποτελεσμάτων, όπως: trees, regression graphs, clustering, sequence modeling, dependency, line analysis κ.λ.π.
7. **Αξιολόγηση των αποτελεσμάτων.** Στο στάδιο αυτό αξιολογούνται τα τελικά αποτελέσματα και συνάγονται συμπεράσματα για τις τάσεις και τις συσχετίσεις που διαφαίνονται στη βάση δεδομένων. Σχετική έκθεση των αποτελεσμάτων διανέμεται σε όλα τα ενδιαφερόμενα μέλη.

Γενικές παρατηρήσεις:

- Τα βήματα αυτά δεν είναι ακολουθιακά. Από οποιοδήποτε στάδιο βρισκόμαστε μπορούμε να επιστρέψουμε σε προηγούμενο να κάνουμε διορθώσεις και να συνεχίσουμε με τα νέα στοιχεία.
- Δεν είναι υποχρεωτικό να ακολουθούνται όλα τα στάδια. Για παράδειγμα, υπάρχουν τρία στάδια όπου γίνεται data cleaning, αν τα δεδομένα είναι όλα χρήσιμα και δεν περιέχουν άχρηστες πληροφορίες που θα καθυστερούσαν το σύστημα θα μπορούσαν να παραληφθούν. Έτσι, θα μπορούσαμε να εφαρμόσουμε αλγόριθμους Data Mining απευθείας σε κάποιο πίνακα της βάσης δεδομένων.

Η ακόλουθη εικόνα παρουσιάζει γραφικά τη διαδικασία που μόλις αναφέρθηκε:



(Περιοδικό BYTE Oct 95)

1.4 Αλγόριθμοι Εξόρυξης Γνώσης

Οι αλγόριθμοι που χρησιμοποιούνται στην καρδιά της KDD διαδικασίας δηλ. στο Data Mining διαχωρίζονται γενικότερα σε 2 κατηγορίες:

1. Verification,
2. Discovery.

Οι αλγόριθμοι της πρώτης κατηγορίας αρκούνται στην **επιβεβαίωση** ερωτήσεων που γίνονται από το χρήστη. Οι ερωτήσεις αυτές έχουν την μορφή υποθέσεων και ο χρήστης περιμένει μια καταφατική ή αρνητική απάντηση από το σύστημα. Στη δεύτερη κατηγορία ανήκουν οι αλγόριθμοι που σκοπό έχουν να **ανακαλύψουν** μοντέλα τα οποία είτε προσδιορίζουν την σημερινή πραγματικότητα είτε προβλέπουν τάσεις για το μέλλον.

Ακολουθεί μια συνοπτική αναφορά σε μερικούς από τους αλγόριθμους που χρησιμοποιούνται για Data Mining:

- **Νευρωνικά δίκτυα:** είναι μη γραμμικά μοντέλα πρόβλεψης. Αρχικά εκπαιδεύονται πάνω στα δεδομένα και μετά χρησιμοποιούνται για την ταξινόμηση - επεξεργασία αυτών.
- **Δένδρα αποφάσεων:** Δενδροειδής δομές που χρησιμοποιούνται για την αναπαράσταση ομάδων δεδομένων. Αυτές οι ομάδες έχουν δημιουργηθεί από την κατηγοριοποίηση των δεδομένων. Υπάρχουν δύο τρόποι κατηγοριοποίησης: Ο ένας έχει να κάνει με **classification** δηλ. ο αναλυτής προσδιορίζει τα κριτήρια που ορίζουν τις περιοχές κατηγοριοποίησης και στη συνέχεια ο αλγόριθμος εντάσσει τα δεδομένα σ' αυτές τις κατηγορίες, ενώ ο άλλος αφορά το **clustering** όπου ο αλγόριθμος είναι υπεύθυνος για την κατηγοριοποίηση των δεδομένων χρησιμοποιώντας διάφορες τεχνικές. Παραδείγματα τέτοιων μεθόδων είναι: classification and regression trees (CART), chi square automatic interaction detection (CHAID).
- **Γενετικοί αλγόριθμοι:** που είναι τεχνικές βελτιστοποίησης. Χρησιμοποιούν διεργασίες όπως η γενετική συνδυαστική, η μετάλλαξη και η φυσική επιλογή, σε ένα σύστημα το οποίο στηρίζεται στην ιδέα της μετάλλαξης.
- **Μέθοδος του κοντινότερου γείτονα:** ταξινομεί τις τιμές μιας στήλης σε ομάδες οι οποίες δημιουργούνται με βάση τις Κ προηγούμενες τιμές των εγγραφών που έχουν αναγνώσει.
- **Εξαγωγή κανόνων :** με χρήση της τεχνικής if-then-else



1.5 Τεχνικές Πρόβλεψης vs Τεχνικών Επιβεβαίωσης

Όπως αναφέρθηκε, οι τεχνικές Data Mining διαχωρίζονται σε δύο τομείς. Ο ένας άφορά την επιβεβαίωση κάποιων υποθέσεων, ενώ ο άλλος αφορά την εξόρυξη-ανακάλυψη και παρουσίαση τάσεων και ενδείξεων.

Πιο συγκεκριμένα, υπάρχουν εργαλεία τα οποία δέχονται ερωτήσεις από τον χρήστη με σκοπό να επιβεβαιώσουν ή να απορρίψουν κάποιες υποθέσεις που έχει κάνει ο χρήστης. Δηλ. δίνουν τη δυνατότητα στον χρήστη (δεν αναφερόμαστε σε end-users αλλά σε ανθρώπους της διοίκησης και του marketing) να κάνει ερωτήσεις σε κάποια δεδομένα, για τα οποία διαφαίνεται (ή δεν διαφαίνεται) μια τάση, με σκοπό τα αποτελέσματα που θα πάρει να διασαφηνίζουν την ύπαρξη ή όχι της τάσης αυτής, ύστερα από σχετική επεξεργασία. Αυτές οι τεχνικές ονομάζονται **τεχνικές επιβεβαίωσης**.

Εφαρμογή της τεχνικής αυτής μπορούν να θεωρηθούν και τα συστήματα OLAP. Τα συστήματα αυτά παρέχουν αυτοματοποιημένες διαδικασίες για Drill Down σε πολυδιάστατες βάσεις δεδομένων. Ένα παράδειγμα είναι το ακόλουθο: ας υποθέσουμε ότι έχουμε ένα αρχείο που περιέχει τις πωλήσεις μιας εταιρείας ανά γεωγραφική περιοχή και πωλητή. Θέλουμε να δούμε αν οι πωλήσεις σε κάποια περιοχή της Ελλάδας παρουσιάζουν κάποιες ιδιαιτερότητες. Η ερώτηση που θα κάναμε (αφού έχουμε προσδιορίσει πλέον το σενάριο που θα ερευνήσουμε) είναι οι αθροιστικές πωλήσεις ανά γεωγραφική περιοχή. Μετά την εξέταση των αποτελεσμάτων θα μπορούσαμε να εντρυφήσουμε σε ένα επίπεδο πιο κάτω για τις περιοχές που παρουσιάζουν μη επιθυμητά αποτελέσματα. Στη συνέχεια αφού βρούμε ποιες είναι αυτές οι περιοχές μπορούμε να ρωτήσουμε ποια είναι τα χαρακτηριστικά των περιοχών αυτών, προκειμένου να βρούμε τις πιθανές αιτίες που οδήγησαν στις επιτυχημένες ή αποτυχημένες πωλήσεις στην περιοχή αυτή.

Συνεπώς, παρατηρούμε ότι η τεχνική επιβεβαίωσης προϋποθέτει δημιουργία πιθανών σεναρίων και στη συνέχεια επεξεργασία αυτών μέχρις ότου μπορέσουμε να απαντήσουμε θετικά ή αρνητικά στο σενάριο που θέσαμε.

Χρησιμοποιώντας αυτά τα εργαλεία ο αναλυτής είναι υποχρεωμένος να σκεφτεί κάποια σενάρια και στη συνέχεια να ελένξει τη συνέπεια τους, με αποτέλεσμα να μην μπορεί να βρει υποδείγματα-ενδείξεις (patterns) σε σημεία που δεν έχει σκεφτεί. Στο σημείο αυτό τη λύση δίνουν οι τεχνικές πρόβλεψης.

Πιο συγκεκριμένα, στην τεχνική αυτή χρησιμοποιούνται μεθοδολογίες οι οποίες εντοπίζουν γεγονότα-καταστάσεις (patterns) που παρουσιάζουν μεγάλη συχνότητα, εντοπίζουν τάσεις και παράγουν συμπεράσματα για τις ενδείξεις που διαφαίνονται από τα περιεχόμενα της βάσης. Τα συμπεράσματα αυτά εξάγονται με τη μικρότερη δυνατή εμπλοκή του χρήστη. Δηλ. Ο χρήστης δεν προσδιορίζει σενάρια, αλλά το σύστημα είναι υπεύθυνο για την εύρεση αυτών, την επεξεργασία τους και την εξαγωγή των αποτελεσμάτων.



Το μειονέκτημα αυτών των τεχνικών είναι ότι παράγουν μεγάλο αριθμό από διαφορετικές ενδείξεις-υποδείγματα (patterns), από τα οποία ο χρήστης πρέπει να διαλέξει αυτά τα οποία είναι χρήσιμα και έχουν νόημα γι' αυτόν.

Συγκρίνοντας τις δύο τεχνικές, συμπεραίνουμε ότι η 'τεχνική επιβεβαίωσης' αναγκάζει τον χρήστη να θέσει σενάρια και να σκεφτεί ο ίδιος που θα μπορούσε να κρύβεται χρήσιμη πληροφορία, ενώ η 'τεχνική πρόβλεψης' αναλαμβάνει να βρει όλες εκείνες τις πληροφορίες-ενδείξεις που θα μπορούσαν να είναι χρήσιμες στον χρήστη, ο οποίος καλείται να διαλέξει τις πραγματικά ενδιαφέρουσες. Τέλος, οι τεχνικές πρόβλεψης είναι σχετικά επικίνδυνες γιατί ενδέχεται να δημιουργήσουν τάσεις που να μην είναι σωστές. Συνεπώς, χρειάζεται καλύτερη επεξεργασία των αποτελεσμάτων.



1.6 Προβλήματα & Δυσκολίες Ανάπτυξης Συστημάτων Εξόρυξης Γνώσης

Ακολουθεί μια λίστα με τις δυσκολίες και τα προβλήματα που είναι δυνατόν να παρουσιαστούν σήμερα κατά τη διαδικασία ανάπτυξης και χρήσης ενός συστήματος KDD:

1. Ανεπαρκή υποστήριξη από τα εργαλεία. Τα περισσότερα εργαλεία για Data Mining δεν καλύπτουν όλους τους αλγόριθμους ενώ μερικά από αυτά δίνουν τη δυνατότητα μόνο για προβλέψεις ή μόνο για διαγνώσεις. Γενικότερα είναι δύσκολο να βρεις εργαλείο που να ενσωματώνει όλες τις δυνατότητες για Data Mining.
2. Πολλές φορές, τα δεδομένα που χρησιμοποιούνται βρίσκονται διεσπαρμένα σε διαφορετικά υπολογιστικά συστήματα της επιχείρησης και σε διαφορετικές πλατφόρμες. Σ' αυτήν την περίπτωση πρέπει να δοθεί μεγάλη έμφαση στο τρίτο στάδιο της διαδικασίας KDD όπου γινόταν το 'data cleaning and preprocessing'. Επίσης, όπως θα αναφέρουμε και αργότερα, σημαντική βοήθεια στο συγκεκριμένο πρόβλημα θα πρόσφερε η δημιουργία ενός Data Warehouse.
3. Η στατιστική εξέταση των δεδομένων μπορεί να έχει σαν αποτέλεσμα τη δημιουργία πολλών μοντέλων πολλά από τα οποία να είναι άχρηστα και μη κατανοητά. Αυτό διορθώνεται με τον σαφή και σωστό προσδιορισμό των παραμέτρων του συστήματος. Χρήση κανόνων μπορεί να συγκεκριμενοποιήσει ακόμη περισσότερο τη λύση.
4. Η δομή των δεδομένων καθώς και η συχνότητα αλλαγής – εισαγωγής δεδομένων στη βάση πρέπει να ληφθεί υπόψη κατά την επιλογή του αλγορίθμου. Για παράδειγμα τα δεδομένα που έχουν να κάνουν με μια βάση που διαχειρίζεται τη διακύμανση του stock αλλάζουν με γρήγορους ρυθμούς, ενώ τα δεδομένα που έχουν σχέση με γεωγραφική θέση αντικειμένων πρέπει να λάβουν ειδικής μεταχείρισης λόγω της ειδικής χωρικής έννοιας που προσδίδουν.



1.7 Data Mining & Data Warehouse

Όπως ήδη αναφέρθηκε, η τεχνολογία Data Warehouse και οι τεχνικές OLAP βρίσκονται ένα σκαλί πίσω από το Data Mining τόσο χρονικά όσο και από πλευράς πληροφορίας που παρουσιάζουν. Το Data Warehouse εμφανίστηκε όταν άρχισαν οι πρώτες ανάγκες για καλύτερη και ταχύτερη πληροφόρηση από τα δεδομένα της βάσης δεδομένων.

Στην πραγματικότητα, το Data Warehouse δεν είναι παρά μια λογική οργάνωση μιας ή πολλών ετερογενών βάσεων δεδομένων με σκοπό την παροχή πιο έξυπνης και ταχείας (on line) αναλυτικής πληροφορίας. Τα συστήματα αυτά έχουν τη δυνατότητα δημιουργίας πολυδιάστατων βάσεων δεδομένων. Αυτό στην πράξη σημαίνει δυνατότητα να ορίσω πολλές μεταβλητές-παραμέτρους στη βάση μου και να ελένξω τις τιμές που παίρνουν κάποιες άλλες μεταβλητές παρακολούθησης μεταβάλλοντας συνεχώς και προς οποιαδήποτε κατεύθυνση το ενδιαφέρον μου.

Συγκεκριμένα, μπορώ να ορίσω σαν ανεξάρτητες μεταβλητές, τη γεωγραφική διάσταση (ιεραρχία=χώρα, νομός, πόλη), τον χρόνο (ιεραρχία=ετος, εξάμηνο, μήνας, ημέρα) και το τμήμα πωλήσεων (όχι ιεραρχία) και να ζητώ τις πωλήσεις που είχα τροποποιώντας δυναμικά τον χρόνο, το γεωγραφικό τμήμα και το τμήμα πωλήσεων, προκειμένου να εξετάσω τις διακυμάνσεις των πωλήσεων.

Το αρνητικό χαρακτηριστικό αυτών των διαδικασιών είναι ότι πρέπει ο αναλυτής να σχεδιάσει στο μιαλό του το σενάριο και στη συνέχεια να ερευνήσει για να διαπιστώσει αν είχε δίκιο ή όχι.

1.7.1 Ορισμός του Data Warehouse

Ακολουθεί ένας ορισμός του data Warehouse:

Data Warehouse είναι μια τεχνική, υποκειμενοστραφής (subject oriented), ολοκληρωμένη (Integrated), Χρονικά εξαρτώμενη (time variant), στατική (volatile), για οργανωμένη συγκέντρωση μεγάλων όγκων δεδομένων με σκοπό την λήψη αποφάσεων (DSS).

Υποκειμενοστραφής: Ο τρόπος με τον οποίο τα δεδομένα θα οργανωθούν, η πληροφορία που θα παρέχεται προσδιορίζεται κάθε φορά από την εκάστοτε επιχείρηση, από τις ανάγκες των αναλυτών της επιχείρησης, από τις ανάγκες του τμήματος marketing, από το είδος της επιχείρησης κ.λ.π.

Ολοκληρωμένη:

Το data warehouse έχει σαν σκοπό την λογική οργάνωση πολλών ετερογενών πηγών δεδομένων. Συνεπώς, είναι σημαντικός στόχος του να μπορεί να διαχειρίζεται και να παρουσιάζει την πληροφορία ικανοποιητικά, γρήγορα και αποτελεσματικά, απ' όποια πηγή κι αν προέρχεται

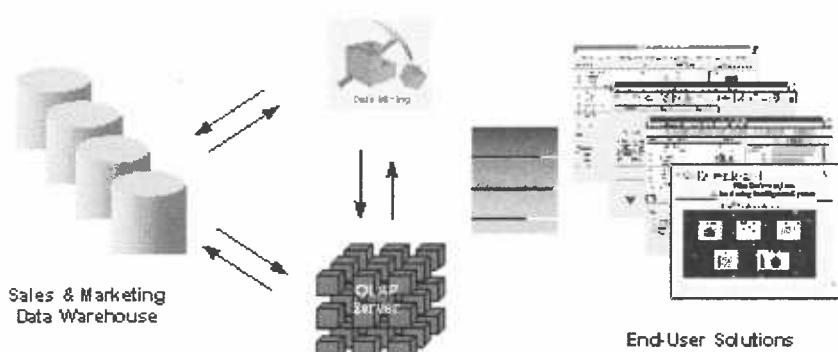


Χρονικά εξαρτώμενη: Στην πραγματικότητα τα δεδομένα που εισάγονται σε ένα data warehouse δεν αλλάζουν. Αποτελούν την καταγραφή της κατάστασης για την συγκεκριμένη χρονική περίοδο στην οποία εισήχθησαν στην βάση. Συνεπώς ένα data warehouse αποτελεί μια συλλογή ιστορικών στοιχείων και αυτό εκφράζει ο όρος χρονικά εξαρτώμενο.

Στατική:

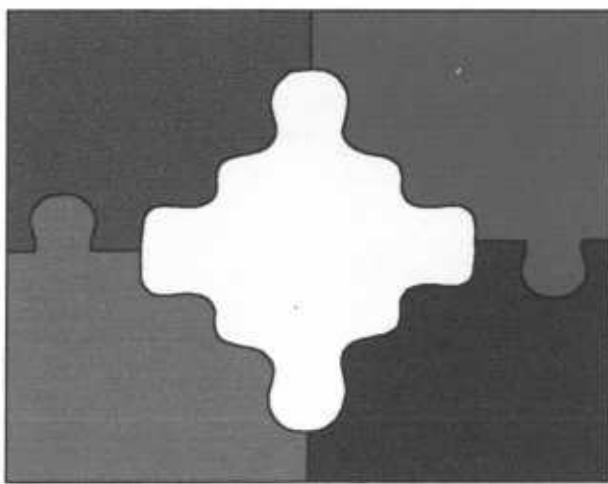
Το data warehouse **δεν** είναι ένα σύστημα συναλλαγών. Αντίθετα, δέχεται πληροφορίες-δεδομένα, τα οργανώνει και στη συνέχεια αναλαμβάνει την παρουσίασή τους. Δεν τα τροποποιεί. Στην χειρότερη περίπτωση υπάρχει η ικανότητα να ξαναφορτωθούν τα δεδομένα κάποιας ημερομηνίας αν αυτά αλλάξουν, αλλά ποτέ δεν τροποποιούνται μέσα στο data Warehouse.

Σε πρακτικό επίπεδο οι τεχνολογίες που αναφέρθηκαν σ' αυτό το κεφάλαιο συνδυάζονται και λειτουργούν από κοινού. Δηλ. υπάρχουν οι βάσεις δεδομένων, οι οποίες ενοποιούνται με την τεχνολογία data warehouse και στη συνέχεια έρχεται το Data Mining για την εξόρυξη της πολύτιμης πληροφορίας-γνώσης. Το σχήμα που ακολουθεί είναι ένα παράδειγμα τέτοιας αρχιτεκτονικής, ωστόσο υπάρχουν πολλά διαφορετικά σχήματα που εφαρμόζονται.



ΚΕΦΑΛΑΙΟ 2^ο

**Τεχνικές Οπτικής Παρουσίασης των
Μοντέλων Data Mining**



Κεφάλαιο 2: Τεχνικές Οπτικής Παρουσίασης των Μοντέλων Data Mining

2.1 Γιατί Χρησιμοποιούμε Τεχνικές Οπτικής Παρουσίασης

Η γραφική παρουσίαση των αποτελεσμάτων των μοντέλων Data Mining έχει σκοπό την ευκολότερη κατανόηση των αποτελεσμάτων. Η διαδικασία KDD εμπλέκει μοντέλα για εξόρυξη άγνωστης πληροφορίας από τις βάσεις δεδομένων και αυτό είναι που κάνει τα αποτελέσματα πολύπλοκα. Για παράδειγμα, αν η πληροφορία, την οποία θα έπαιρνε ο αναλυτής είχε να κάνει με τα επίπεδα των πωλήσεων κάποιου προϊόντος, τότε δεν θα δυσκολευόταν να την καταλάβει.

Αν, όμως, το αποτέλεσμα της διαδικασίας έχει να κάνει με συσχετισμένη πληροφορόρηση μεταξύ γνωστών και μη παραγόντων, τότε η μελέτη των αποτελεσμάτων και η απόδοσή τους σε επιχειρησιακά δεδομένα εμπεριέχει μεγάλο βαθμό δυσκολίας. Τελικά, η οπτική παρουσίαση των αποτελεσμάτων μειώνει το χρόνο κατανόησής τους και κάνει πιο απλή αυτή τη διαδικασία.

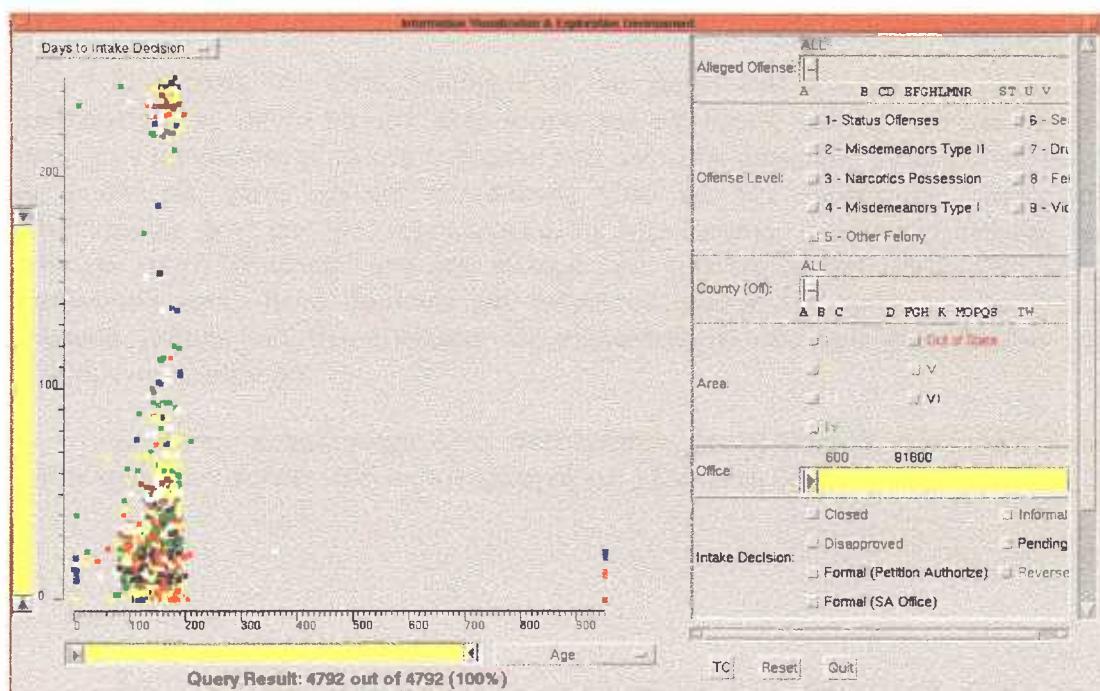
Ένας άλλος λόγος οπτικής παρουσίασης των αποτελεσμάτων είναι η εμπιστοσύνη του χρήστη. Για παράδειγμα, όταν ο project manager αποφασίσει, ύστερα από μελέτη των αποτελεσμάτων του Data Mining, ότι χρειάζεται 3 μήνες για την ολοκλήρωση ενός έργου, τότε θα είναι ευκολότερο γι' αυτόν να αποκτήσει την εμπιστοσύνη του τοπ manager, εφόσον τα αποτελέσματα του Data Mining είναι ευπαρουσίαστα και ευκατανόητα.

Συμπερασματικά, καταλήγουμε στο ότι ένα εργαλείο Data Mining θα πρέπει να έχει τη δυνατότητα της γραφικής παρουσίασης των αποτελεσμάτων της έρευνας των δεδομένων και θα πρέπει σ' ένα δεύτερο επίπεδο να επιτρέπει στο χρήστη τη δυναμική επίδραση πάνω στο γράφημα, με σκοπό την απάντηση απλών ερωτήσεων ή την προβολή στατιστικών στοιχείων μεγαλύτερης λεπτομέρειας.

Βέβαια, η λογική αυτή εμπεριέχει παγίδες οι οποίες μπορούν να μετατρέψουν τις τεχνικές οπτικής παρουσίασης των αποτελεσμάτων σε μπούμερανγκ. Τα αποτελέσματα που παράγονται από τις διαδικασίες εξόρυξης γνώσης προέρχονται από πολύ μεγάλες βάσεις δεδομένων, και αυτό σημαίνει ότι μπορεί να είναι ιδιαίτερα πολύπλοκα και οι συνδυασμοί που παράγονται να είναι υπερβολικά πολλοί και να έχουν δύσκολα παρουσιάσιμη δομή. Συνεπώς, οι τεχνικές οπτικής παρουσίασης πρέπει να είναι δομημένες σωστά, ώστε να εκφράσουν με απλό και ευνόητο τρόπο τα αποτελέσματα. Όπως είπε και ο Kuhn: "Since human perception imposes an upper bound on the complexity of graphic representation, only a small number of relations can be shown" (Ellson 1990; Kuhn 1990).

Επίσης, μερικές φορές τα αποτελέσματα, τα οποία λαμβάνουμε από τα γραφήματα είναι τέτοια, που μειώνουν την εμπιστοσύνη μας για την αξιοπιστία της διαδικασία Data Mining και για τις τεχνικές οπτικής παρουσίασης. Για παράδειγμα, στο γράφημα 1 παρουσιάζονται: οι ημέρες που χρειάστηκαν κάποια παιδιά για να

δραστηριοποιηθούν σε μια πράξη σε σχέση με την ηλικία των παιδιών αυτών (σε μήνες).



Γράφημα 1

Το αξιοσημείωτο σ' αυτό το γράφημα είναι ότι υπάρχουν μερικές κουκίδες που αναφέρονται σε παιδιά 80 ετών. Ωστόσο, αυτό το γεγονός δεν θα είχε συμβεί αν είχε γίνει αποτελεσματικό και σωστό data cleaning, το οποίο αποτελεί μέρος της διαδικασίας KDD. Συνεπώς, αυτή η παραπλανητική πληροφορία που παρουσιάζεται, δεν οφείλεται ούτε στον αλγόριθμο Data Mining ούτε στις τεχνικές οπτικής παρουσίασης των αποτελεσμάτων αλλά στην παράκαμψη ή στη λάθος ενεργοποίηση ενός βήματος της διαδικασίας KDD (Data Cleaning).

2.2 Αρχές και Κανόνες που Διέπουν τις Τεχνικές Οπτικής Παρουσίασης

Ο σκοπός του Data Mining είναι να δώσει στο χρήστη ένα κατανοητό αποτέλεσμα, το οποίο θα τον πληροφορεί για τις τάσεις και τις συσχετίσεις που υπάρχουν κρυμμένες μέσα στα δεδομένα του. Ωστόσο, όπως ήδη αναφέρθηκε, αυτές οι διαδικασίες είναι σχετικά πολύπλοκες τόσο για την εύρεση της επιθυμητής πληροφορίας όσο και για την παρουσίασή της. Επίσης, η απουσία γνώσης των πιθανών αποτελεσμάτων, από την πλευρά του χρήστη, οδηγεί εύκολα σε λάθος κατανόησή τους. Αυτό σημαίνει ότι πρέπει να αναλυθεί ο τρόπος σκέψης ενός πιθανού χρήστη και στη συνέχεια να σχεδιαστεί το σύστημα στηριζόμενο στις εξαγόμενες παραδοχές.

Συνεπώς, καταλήγουμε στο συμπέρασμα ότι η οπτική παρουσίαση των αποτελεσμάτων πρέπει να είναι **κατανοητή και αξιόπιστη** [18].

- **Κατανοητή.**

Κατανοητό αποτέλεσμα σημαίνει κατανοώ τα αποτελέσματα σε σχέση με το πλαίσιο των δεδομένων που μελετώ και σε σχέση με το πρόγραμμα που χρησιμοποιώ. Το κάθε πρόγραμμα χρησιμοποιεί αλγορίθμους που δίνουν βάρος σε διαφορετικά σημεία και κατά συνέπεια η άγνοια (από την πλευρά του χρήστη) της λογικής του προγράμματος μπορεί να οδηγήσει σε λανθασμένες αποφάσεις. Συνεπώς, η παρουσίαση των αποτελεσμάτων πρέπει να ελαχιστοποιεί τέτοιες πιθανότητες, καθώς άλλωστε οι τεχνικές οπτικής παρουσίασης των αποτελεσμάτων Data Mining δεν απευθύνονται μόνο σε στατιστικούς αλλά και σε απλούς χρήστες (π.χ. managers).

Μεγάλη σημασία έχει επίσης και η δυνατότητα αλληλεπίδρασης με τα αποτελέσματα ώστε να μπορεί ο χρήστης να δίνει απάντηση σε μικρές αλλά κρίσιμες ερωτήσεις που μπορεί να δημιουργηθούν κατά τη διαδικασία της εξερεύνησης, π.χ. έστω ένας χρήστης που παρατηρεί τις πωλήσεις της εταιρείας του σε επίπεδο νομών της Ελλάδας, και διαπιστώνει μια συσχέτιση μεταξύ δύο νομών. Ως άμεσο αποτέλεσμα θέλει να δει με ποιο τρόπο αναλύονται οι πωλήσεις σε κάθε έναν ξεχωριστά. Αν δεν έχει τη δυνατότητα να κάνει drill down τότε πρέπει να γυρίσει στο προηγούμενο μενού και να θέσει την ίδια ερώτηση προσθέτοντας το νομό που επιθυμεί να δει. Υποθέτοντας ότι αυτή η διαδικασία θα επαναληφθεί συχνά και με δεδομένο ότι η δυνατότητα του ανθρώπου να αποθηκεύει προσωρινά γνώση στον εγκέφαλό του περιορίζεται σε 5-7 στοιχεία-αντικείμενα ανά πάσα στιγμή, καταλαβαίνουμε ότι μετά από 3-4 ερωτήσεις θα έχει ξεχάσει τι έψαχνε. Είναι δύσκολο, ύστερα από μια μεγάλη διαδικασία ερωτήσεων, να κρατήσει στο μυαλό του τη λογική συνέχεια αυτών για να βγάλει συμπεράσματα.

Τρεις παράγοντες προσδιορίζουν τον βαθμό κατανόησης των αποτελεσμάτων:

1. **Παρουσίαση:** πρέπει να δοθεί μεγάλη βαρύτητα στην *ισοστάθμιση* των παραγόντων: πολυπλοκότητα/απώλεια στοιχείων. Το γράφημα θα πρέπει να

είναι απλό ώστε να συμβαδίζει με τον τρόπο σκέψης του χρήστη καθώς και να μην χάνει σε ποιότητα.

2. **Αλληλεπίδραση:** Ο χρήστης πρέπει να έχει τη δυνατότητα επικοινωνίας με το οπτικό αποτέλεσμα. Ήα πρέπει να λειτουργεί μια διαδικασία επανατροφοδότησης στοιχείων (από το χρήστη προς το σύστημα) ώστε μια νέα ερώτηση να είναι επέκταση της παλιάς, και μάλιστα αυτό να γίνεται άμεσα (π.χ. Drill Down, Roll Up, Slice & Dice κ.λ.π.)
3. **Ολοκλήρωση:** Να μπορεί το οπτικό αποτέλεσμα να δώσει μια ολοκληρωμένη εικόνα στον χρήστη για τα δεδομένα του.

- **Αξιόπιστη.**

Σημαντικός παράγοντας δημιουργίας εμπιστοσύνης για ένα μοντέλο από την πλευρά του χρήστη είναι ο παράγοντας "Χρόνος χρήσης". Ο κάθε χρήστης θα μπορεί να θεωρήσει ότι το μοντέλο που χρησιμοποιεί είναι αξιόπιστο μόνο όταν για μεγάλο χρονικό διάστημα χρήσης δέχεται σωστά αποτελέσματα.

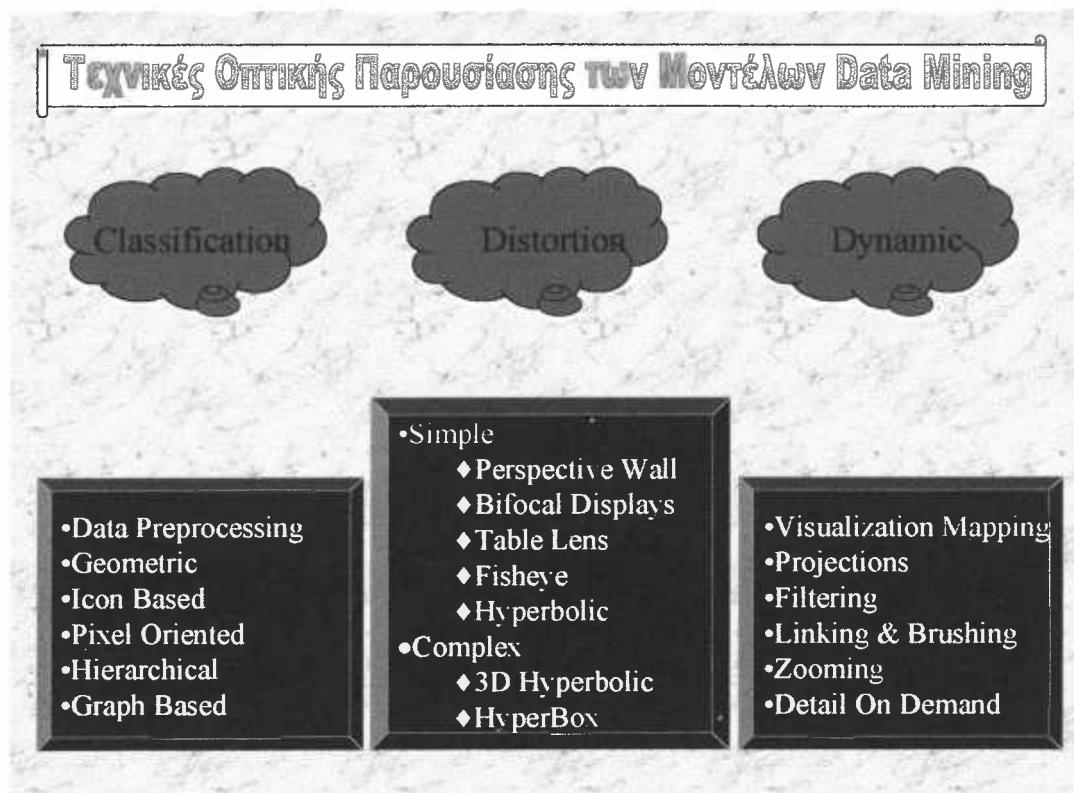
Ωστόσο υπάρχουν διάφοροι άλλοι παράγοντες, οι οποίοι μπορούν να πείσουν για την αξιοπιστία του μοντέλου.

1. Είναι απαραίτητο το μοντέλο να δηλώνει με σαφήνεια μέσω της οπτικής παρουσίασης το βαθμό εμπιστοσύνης των αποτελεσμάτων. Κάθε στατιστική διαδικασία ανεύρεσης συσχετίσεων ή τάσεων μέσα στα δεδομένα εξάγει ένα σύνολο αποτελεσμάτων τα οποία χαρακτηρίζονται από ένα ποσοστό εμπιστοσύνης.
2. Το κάθε πρόγραμμα έχει κάποιους περιορισμούς, τους οποίους δεν πρέπει να τους κρύβει από τον χρήστη, αλλά να τους καθορίζει διακριτικά, ώστε ο χρήστης να μην ξεπερνάει τα όρια χρήσης του, οδηγώντάς το σε ανακριβή αποτελέσματα.
3. Η παρουσίαση των αποτελεσμάτων θα πρέπει να είναι σαφής και κατανοητή αποφεύγοντας σύνθετα αποτελέσματα. Πολλές φορές ένα δισδιάστατο ραβδόγραμμα αποδεικνύεται καλύτερο από ένα πολυδιάστατο γράφημα.
4. Το αποτέλεσμα πρέπει να πλησιάζει τον τρόπο σκέψης του χρήστη.



2.3 Τεχνικές Οπτικής Παρουσίασης

Στο κεφάλαιο αυτό θα γίνει μια αναλυτική παρουσίαση των υπαρχόντων τεχνικών οπτικής παρουσίασης των αποτελεσμάτων των διαδικασιών Data Mining. Οι τεχνικές αυτές είναι ειδικά σχεδιασμένες για να μπορούν να παρουσιάζουν μεγάλους όγκους δεδομένων και απευθύνονται κυρίως σε διαδικασίες Data Mining. Ακολουθεί ένα διάγραμμα, το οποίο κατηγοριοποιεί τις τεχνικές αυτές διαχωρίζοντάς τις σε ομάδες ανάλογα με τα χαρακτηριστικά τους.



Πίνακας κατηγοριών τεχνικών οπτικής παρουσίασης μοντέλων Data Mining [1]

Στις επόμενες παραγράφους θα αναλυθούν πιο λεπτομερώς οι τεχνικές αυτές.

2.3.1 Τεχνικές Μείωσης Διαστάσεων

Μια κατηγορία τεχνικών που τοποθετούνται στην κατηγορία Classification είναι οι τεχνικές "Data Preprocessing". Σκοπός των τεχνικών αυτών είναι η παρουσίαση k-διάστατων δεδομένων σε δ-διάστατο χώρο, όπου $d < k$ για εύκολη αναπαράστασή τους. Μια πρόσφατη προσπάθεια σε τέτοιου είδους τεχνικές είναι ο αλγόριθμος FastMap [23].

Βασικός στόχος του αλγόριθμου FastMap είναι η αντιστοίχηση των δεδομένων σε σημεία χρησιμοποιώντας την πληροφορία της ομοιογένειας και της απόστασης των δεδομένων μεταξύ τους.

Ορισμοί:

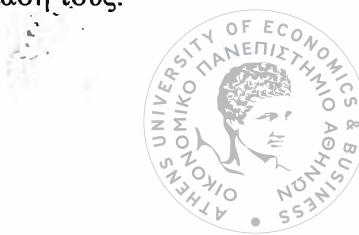
- Το K-διάστατο σημείο P που παρουσιάζει το αντικείμενο (της βάσης) O_i ονομάζεται εικόνα του O_i , $P_i = (\chi_{i1}, \chi_{i2}, \chi_{i3}, \chi_{i4}, \dots)$.
- Ο χώρος απεικόνισης όλων των εικόνων ονομάζεται target space.
- Query by example είναι οι ερωτήσεις που τίθενται προκειμένου να βρεθούν x στοιχεία που απέχουν μια συγκεκριμένη (ορισμένη από το χρήστη) απόσταση σχετικά με την τεθείσα ερώτηση.
- All pairs ονομάζει την ερώτηση, η οποία έχει ως αποτέλεσμα την συγκέντρωση όλων των ζευγαριών των αντικειμένων που απέχουν μια συγκεκριμένη (ορισμένη από το χρήστη) απόσταση μεταξύ τους (ή και μικρότερη απόσταση από την ορισμένη).

Πρόβλημα προς λύση:

- Distance Case: Δεδομένων N αντικειμένων και των αποστάσεων μεταξύ τους, βρες N σημεία στον K διάστατο χώρο στα οποία οι αποστάσεις να είναι ορισμένες με τον καλύτερο δυνατό τρόπο. Δηλ. αν υποτεθεί ότι είναι δεδομένες οι αποστάσεις που ορίζονται μεταξύ των δεδομένων τότε ζητείται να βρεθούν τα σημεία εκείνα στον k-διάστατο χώρο που παρουσιάζουν με τη μικρότερη δυνατή αλλοίωση τις αποστάσεις αυτές.

Πλεονεκτήματα της τεχνικής αυτής είναι:

1. Μειώνει το χρόνο αναζήτησης για ερωτήσεις του τύπου: βρες τις εγγραφές που ανήκουν σε ένα συγκεκριμένο εύρος τιμών.
2. Βοηθάει με την οπτική παρουσίαση την εύρεση των clusters στις διαδικασίες Data Mining.
3. Χρησιμοποιεί τις αποστάσεις των δεδομένων για την παρουσίασή τους, συνεπώς δεν χρειάζεται κανονικοποίηση για την ορθή παρουσίασή τους.



Άλλες μέθοδοι:

Στο σημείο αυτό θα αναφερθούν περιληπτικά 2 τεχνικές που έχουν τον ίδιο στόχο με τη FastMap. Ο λόγος που αναφέρονται είναι για να εξαχθούν συγκριτικά αποτελέσματα, και για να παρουσιαστούν κάποιοι μαθηματικοί τύποι που χρησιμοποιούνται από τον FastMap αλγόριθμο.

MDS (Multi Dimensional Scaling)

Ο αλγόριθμος αυτός χρησιμοποιείται με σκοπό την εύρεση της υπάρχουνσας δομής ενός συνόλου δεδομένων βασιζόμενος στην ομοιότητα (similarity) και την ανομοιογένεια (dissimilarity) που έχουν τα δεδομένα μεταξύ τους. Απαραίτητα στοιχεία για την εφαρμογή της μεθόδου είναι: τα δεδομένα, οι αποστάσεις τους και η επιθυμητή διάσταση παρουσίασης.

Ορίζεται η μετρική stress = $\text{square}(\sum_{i,j} (\delta_{ij} - d_{ij})^2 / \sum_{i,j} d_{ij}^2)$, όπου δ_{ij} είναι η μετρική ανομοιογένειας μεταξύ των αντικειμένων O_i και O_j , d_{ij} είναι η απόσταση μεταξύ των εικόνων τους P_i και P_j . Η μετρική stress μας δίνει το μέσο όρο του σχετικού λάθους των αποστάσεων, επί του συνόλου των δεδομένων. Ουσιαστικά, η μέθοδος αυτή βρίσκει την απόσταση ενός αντικειμένου προς τα υπόλοιπα $N-1$ αντικείμενα και σταδιακά μειώνει το λάθος που υπεισέρχεται μειώνοντας τη μετρική stress. Η τεχνική αυτή εφαρμόζει τον αλγόριθμο "steepest descent", δηλ. αφού υπολογίζει μια πρώτη προσέγγιση εφαρμόζει επαναληπτικά την διαδικασία μειώνοντας το λάθος.

Απαιτεί χρόνο $O(N^2)$ όπου N είναι ο αριθμός των αντικειμένων. Ένα σημαντικό μειονέκτημα της τεχνικής είναι ότι ο υπολογισμός των αποστάσεων ενός νέο-εισαγόμενου αντικειμένου καταλήγει να είναι πολύ χρονοβόρος ($O(N)$ στην καλύτερη περίπτωση).

Αλγόριθμος Karhunen-Loeve

Ο αλγόριθμος Karhunen-Loeve αναπαριστά σύνολα σημείων που παρουσιάζουν μια συσχέτιση μεταξύ τους με διανύσματα. Για το λόγο αυτό χρησιμοποιείται κυρίως για εύρεση υποδειγμάτων-μοντέλων μέσα στα δεδομένα, δηλ. σύνολα σημείων που μπορεί να συσχετίζονται μεταξύ τους.

Πλεονεκτήματα: μειώνει το mean square λάθος απεικόνισης του αντικειμένου στον K - διάστατο χώρο.

Μειονεκτήματα: Είναι επίσης αργός.

Anάκτηση και Κατηγοριοποίηση (Retrieval και Clustering)

Τρεις κλάσεις ομαδοποιούν όλες τις τεχνικές αυτού του είδους και αυτές είναι:



1. Tree based methods (R-tree, P-tree, B-tree),
2. Methods using linear quadtrees (ή ομοίως z-ordering, space filling curves),
3. Methods using grid-files.

Οι παραπάνω μέθοδοι χρησιμοποιούνται για να μειώσουν το εύρος αναζήτησης σε μια ερώτηση όπου η απάντηση είναι εύρος τιμών. Καμιά από τις μεθόδους αυτές δεν προσπαθεί να αντιστοιχίσει τιμές σε K-διάστατο χώρο.

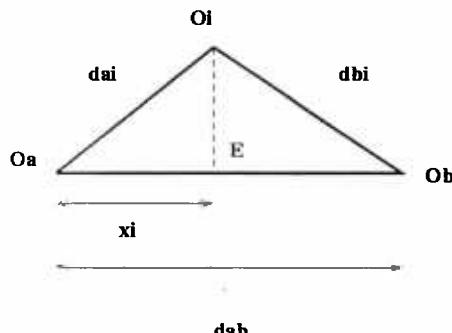
Αλγόριθμος Fast-Map

Η βασική ιδέα της τεχνικής αυτής βασίζεται στο ότι το κάθε στοιχείο από τα δεδομένα της βάσης αποτελεί ένα σημείο στο K-διάστατο χώρο και μπορεί να απεικονιστεί σ' αυτόν υπολογίζοντας τις αποστάσεις του από κάθε διάσταση. Τα δεδομένα που χρειάζεται η τεχνική αυτή είναι (a) ένα σύνολο από N αντικείμενα, (b) μια συνάρτηση υπολογισμού των αποστάσεων $D(O_i, O_j)$ (Πίνακας 1), (c) τον αριθμό των διαστάσεων k.

Ο τρόπος υπολογισμού των αποστάσεων στην k διάσταση έχει σαν βασική ιδέα τον υπολογισμό των αποστάσεων από μια συγκεκριμένη γραμμή, η επιλογή της οποίας είναι πολύ βασική. Για κάθε άλλο αντικείμενο της βάσης παίρνουμε την προβολή του στη γραμμή αυτή. Με βάση τον τύπο

$$\text{Cosine law: } d_{b,i}^2 = d_{a,i}^2 + d_{a,b}^2 - 2x_i d_{a,b} \Rightarrow x_i = (d_{a,i}^2 + d_{a,b}^2 - d_{b,i}^2) / 2d_{a,b}$$

μπορούμε να υπολογίσουμε την απόσταση της προβολής του αντικειμένου από τα δύο σημεία της γραμμής στην οποία το προβάλουμε (Εικόνα 1).

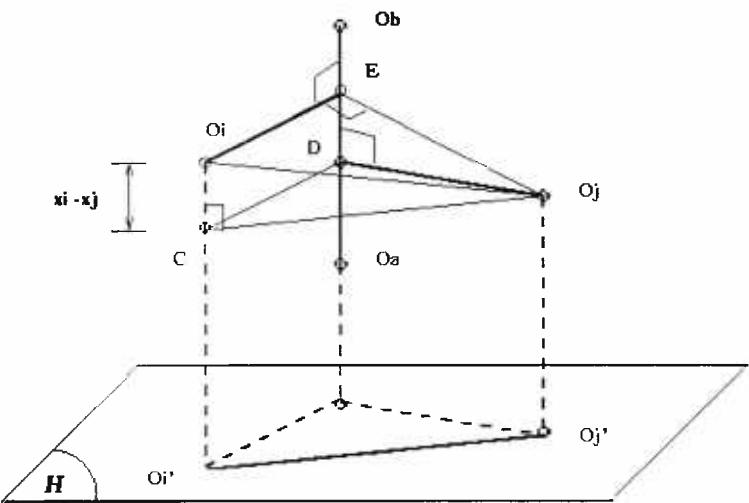


Εικόνα 1

Κατ' ουσία, αυτή είναι η λύση του προβλήματος στον μονοδιάστατο χώρο. Τι συμβαίνει όμως στον πολυδιάστατο; Έστω ένα hyper-plane το οποίο είναι κάθετο στην γραμμή που ορίσαμε πριν, και έστω οι προβολές των αντικειμένων σ' αυτό. Με βάση τον ακόλουθο τύπο

$$(D'(O_i', O_j'))^2 = (D(O_i, O_j))^2 - (x_i - x_j)$$

μπορούμε να υπολογίσουμε τις αποστάσεις των προβολών των αντικειμένων σε ένα δεύτερο επίπεδο πάνω στο hyper-plane (Εικόνα 2). Με αυτόν τον τρόπο λύνεται το πρόβλημα των δύο διαστάσεων. Επαναλαμβάνοντας αυτό το βήμα k φορές λύνεται το πρόβλημα των k διαστάσεων.



Εικόνα 2

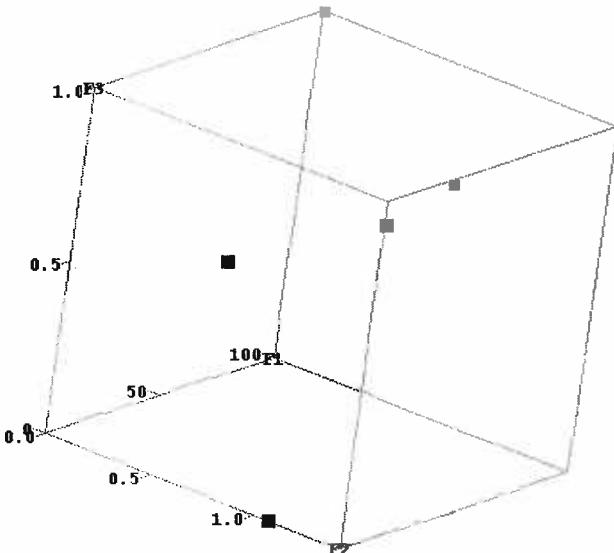
Ένα σημαντικό θέμα για την υλοποίηση των παραπάνω είναι η επιλογή της γραμμής πάνω στην οποία θα γίνονται οι προβολές. Η γραμμή αυτή πρέπει να είναι όσο πιο μεγάλη γίνεται προκειμένου να μην δημιουργείται συνωστισμός από τις προβολές των άλλων αντικειμένων πάνω της. Έτσι, πρέπει να βρεθούν τα δεδομένα O_a και O_b που μεγιστοποιούν τη συνάρτηση $D(O_a, O_b)$ (Πίνακας 3). Δεδομένου ότι έχουμε την πληροφορία των αποστάσεων των αντικειμένων μεταξύ τους χρησιμοποιείται ο ακόλουθος heuristic αλγόριθμος:

1. Διάλεξε τυχαία ένα αντικείμενο από τη βάση και θέσε το να είναι το δεύτερο αντικείμενο που θα ορίζει τη γραμμή (O_b).
2. Θέσε σαν αντικείμενο O_a αυτό που μεγιστοποιεί την συνάρτηση D με βάση το O_b που έχει επιλεγεί.
3. Βρες το αντικείμενο O_b που μεγιστοποιεί τη συνάρτηση D για το επιλεγμένο O_a .
4. Επανέλαβε τα βήματα 2, 3 όσο χρειάζεται.

Ακολουθεί περιγραφή του αλγορίθμου FastMap.

1. Δημιουργησε ένα πίνακα $X[] = N \times k$, όπου N ο αριθμός των αντικειμένων και k οι διαστάσεις. Για κάθε αντικείμενο θα υπολογίζονται οι αποστάσεις του για κάθε διάσταση.
2. Δημιουργησε ένα πίνακα $PA[] = 2 \times k$, όπου θα αποθηκεύονται τα αντικείμενα που χρησιμοποιήθηκαν για το σχηματισμό της γραμμής σε κάθε επίπεδο.
3. Βρες τα αντικείμενα που θα συνθέσουν τη γραμμή O_a, O_b .
4. Για κάθε άλλο αντικείμενο O_i υπολόγισε την απόσταση της προβολής του από τη γραμμή που σχηματίζουν τα δύο ρίνοτα αντικείμενα, δηλ. υπολόγισε τα x_i
5. Ξανακάλεσε τη διαδικασία αυτή για την επόμενη διάσταση.

Το αποτέλεσμα αυτής της διαδικασίας είναι ένας πίνακας $X[]$ ο οποίος περιέχει τις



Εικόνα 3

αποστάσεις του κάθε αντικειμένου από το κάθε επίπεδο (Πίνακας 2). Η Γραφική απεικόνιση των παρουσιάζεται στην Εικόνα 3.

	O1	O2	O3	O4	O5
O1	0	1	1	100	100
O2	1	0	1	100	100
O3	1	1	0	100	100
O4	100	100	100	0	1
O5	100	100	100	1	0

Πίνακας 1: Αποστάσεις Δεδομένων

X[]	f_1	f_2	f_3
O1	0	0.707089	0.668149
O2	0.005	1.414118	0.935411
O3	0.005	1.06062	0
O4	100	0.707089	0.668149
O5	99.995	0	1

Πίνακας 2: Αποτελέσματα FastMap

Iteration #	Pivot	Stress
1	O1, O4	0.008
2	O5, O2	0.004
3	O3, O5	0.001

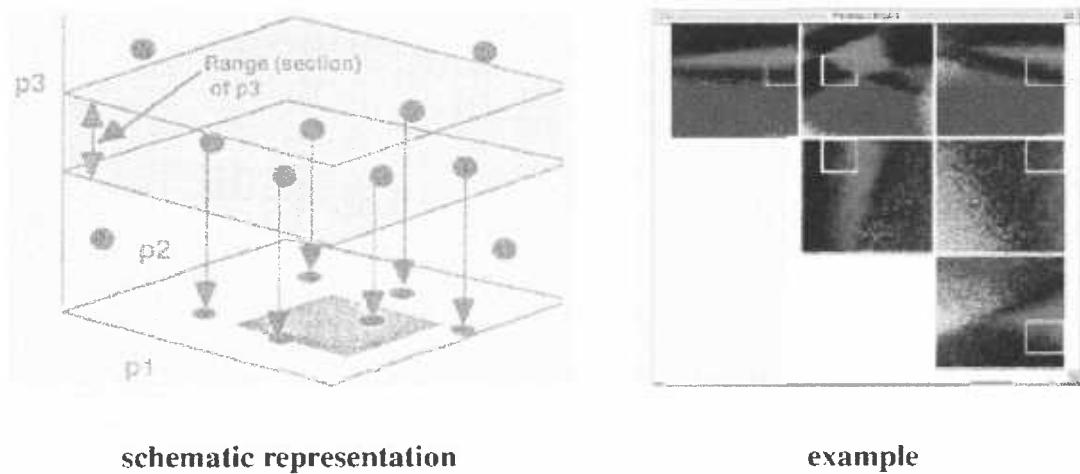
Πίνακας 3: Επιλεγμένα δεδομένα σχηματισμού των γραμμών

2.3.2 Γεωμετρικές Τεχνικές (Geometric)

Οι γεωμετρικές τεχνικές αποσκοπούν στην εύρεση προβολών πολυδιάστατων δεδομένων σε δισδιάστατο χώρο και στην επιλογή εκείνων που δείχνουν ενδιαφέρουσες. Ακολουθεί περιγραφή των τεχνικών αυτών.

Τεχνική Προβολών (Projection Views)

Οι τεχνικές αυτές αναζητούν τις προβολές εκείνες οι οποίες θα δώσουν ένα ενδιαφέρον αποτέλεσμα, ή βοηθούν το χρήστη να προσδιορίσει τις προβολές αυτές[1]. Συγκεκριμένα, επιλέγεται στον k-διάστατο χώρο κάποιο εύρος τιμών τις οποίες θέλουμε να προβάλουμε, και χρωματίζουμε τις προβολές αυτών στο δισδιάστατο χώρο με ένα συγκεκριμένο χρώμα. Παρατηρώντας την κατανομή των τιμών για το συγκεκριμένο εύρος τιμών (πεδίο ορισμού) που επιλέξαμε μπορούμε να εξάγουμε χρήσιμα συμπεράσματα (Εικόνα 4).



schematic representation

example

Εικόνα 4

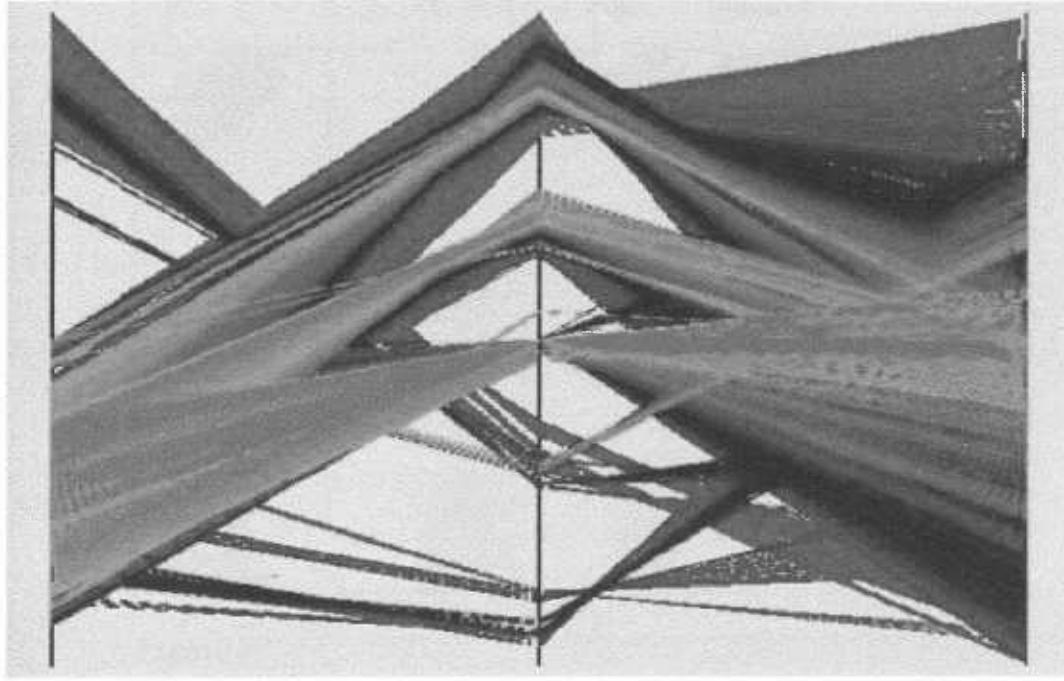
Παράλληλη Γραμμική Αναπαράσταση (Parallel coordinate)

Και αυτή η τεχνική έχει σαν σκοπό να προβάλει τα k-διάστατα δεδομένα στον δισδιάστατο χώρο. Συγκεκριμένα αποτελείται από n ισαπέχοντες άξονες οι οποίοι είναι παράλληλοι σε ένα από τους άξονες (x, y) και οι οποίοι αντιπροσωπεύουν τα n γνωρίσματα ενός συνόλου δεδομένων. Οι άξονες αυτοί είναι ισομήκεις και περιλαμβάνουν όλο το πεδίο τιμών της κάθε μεταβλητής που αντιπροσωπεύουν. Κάθε μονάδα (εγγραφή) από το σύνολο των δεδομένων αντιπροσωπεύεται από μία πολυγωνική γραμμή η οποία τέμνει τον κάθε άξονα στο σημείο που αντιστοιχεί στην τιμή της για το συγκεκριμένο γνώρισμα (Εικόνα 5).

Η παραπάνω τεχνική χρησιμεύει για την διαπίστωση χαρακτηριστικών στα δεδομένα, όπως κατανομή των δεδομένων και συναρτησιακές εξαρτήσεις.

Μειονεκτήματα: Αδυναμία αναπαράστασης μεγάλων όγκων δεδομένων. Σ' αυτή την περίπτωση οι γραμμές που αναπαριστούν τα δεδομένα επικαλύπτονται και υπάρχει η πιθανότητα να δημιουργηθεί ένα δυσανάγνωστο αποτέλεσμα.

Parallel Coordinates



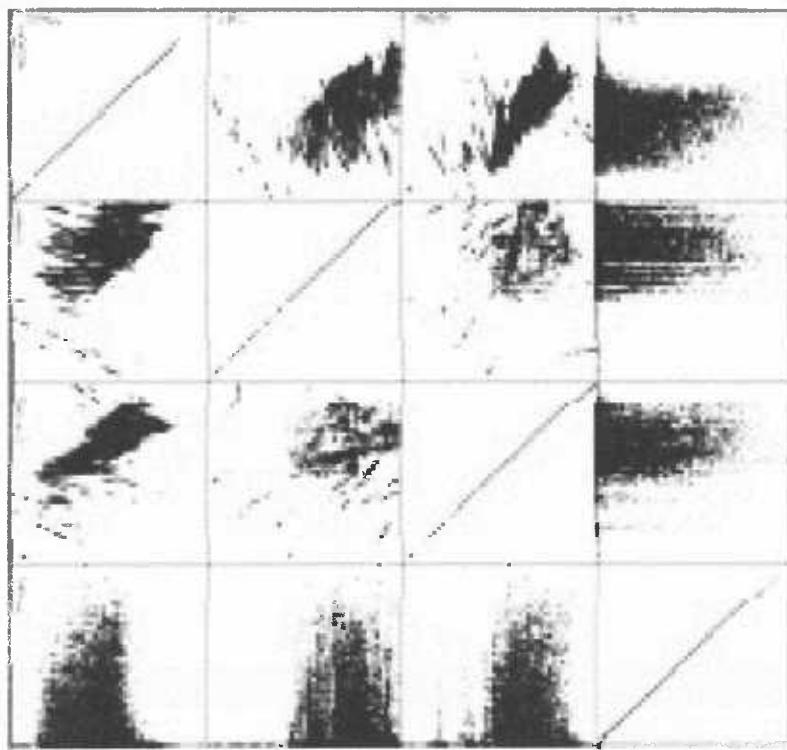
Εικόνα 5

Πίνακες από Scatterplots

Πρόκειται για ένα πίνακα από Scatterplots τα οποία αναπαριστούν τις k διαστάσεις των δεδομένων. Κάθε Scatterplot του πίνακα παρουσιάζει σε δύο άξονες (x,y) την κατανομή των τιμών, για δύο από τις k μεταβλητές. Συνεπώς, δημιουργείται ένας k -διάστατος ($k^2 \cdot k$) πίνακας ο οποίος παρουσιάζει ένα Scatterplot για κάθε συνδυασμό των μεταβλητών που χρησιμοποιούνται (Εικόνα 6).

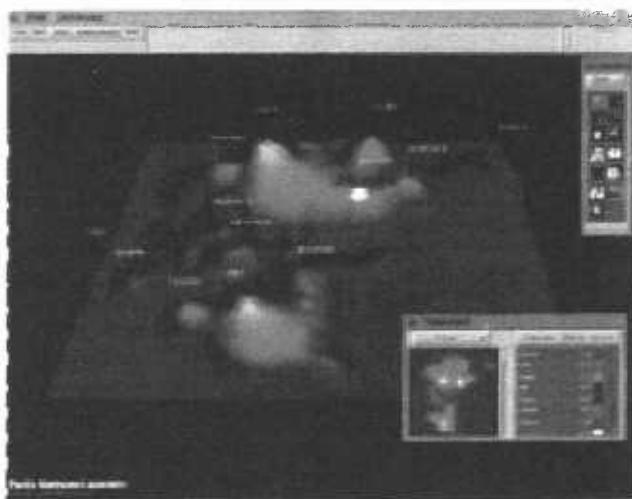
Μειονεκτήματα: Στην ουσία τα μισά διαγράμματα αποτελούν επαναλήψεις των υπολοίπων. Συγκεκριμένα, το διάγραμμα με τις μεταβλητές (x,y) αποτελεί την περιστροφή (κατά 90°) του διαγράμματος με τις μεταβλητές (y,x), και το διάγραμμα με τις μεταβλητές (x,x) δεν παρέχει σημαντική πληροφόρηση. Επομένως, ένα μειονεκτήμα της μεθόδου είναι η σπατάλη του διαθέσιμου χώρου παρουσίασης. Σε κάθε περίπτωση, τα γραφήματα που έχουν αξία είναι $(k^2-k)/2$, δηλ. εκμετάλλευση χώρου οθόνη μικρότερη από το 50% $((k^2-k)/2)/k^2 = [50\% - (1/2k)]$.

Scatterplot-Matrices



Εικόνα 6

Landscapes



Εικόνα 7

Τοπογράφημα (Landscapes)

Τα δεδομένα παρουσιάζονται με προοπτική τοπίου σε τρισδιάστατο χώρο. Στην τεχνική αυτή τα δεδομένα πρέπει να τύχουν επεξεργασίας η οποία θα δημιουργεί την απαραίτητη πληροφορία για την χωρική αναπαράστασή τους. Απαιτείται προσοχή για να μην οδηγηθούμε σε λάθος αποτελέσματα και ταυτόχρονα να μπορέσουμε να βγάλουμε συμπεράσματα για τα χαρακτηριστικά που διακρίνονται στα δεδομένα (Εικόνα 7).

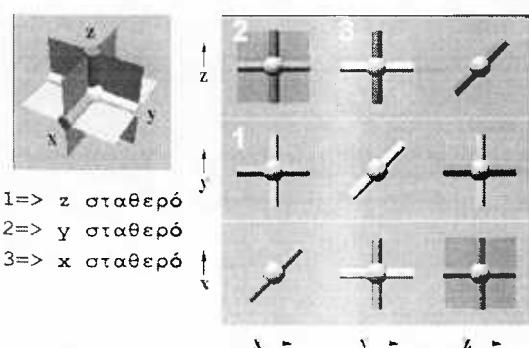
Δισδιάστατα Δείγματα του k Διάστατου Χώρου (Hyperslice)

Το Hyperslice είναι μια εξελιγμένη μορφή του Scatterplot Matrices. Η τεχνική Scatterplot matrices παρουσιάζει πολυδιάστατα δεδομένα σε ένα πίνακα από Scatterplots όπου κάθε ένα από αυτά αφορούσε ένα συνδυασμό μεταβλητών. Αυτή η λογική δεν απεικόνιζε πουθενά όλες τις διαστάσεις - μεταβλητές του συνόλου δεδομένων, άρα δεν θα χαρακτηριζόταν ως καθαρότατη τεχνική.

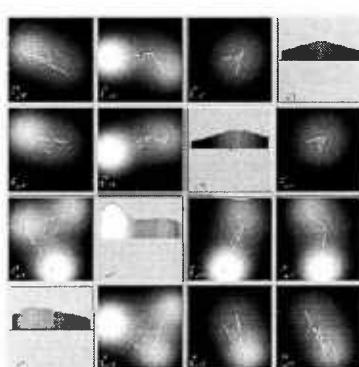
Τα Hyperslices [5][25] χρησιμοποιούν τεχνογνωσία από την Scatterplot Matrices και την Projection Views. Γνωρίζοντας ότι δεν μπορούν πολυδιάστατα δεδομένα να παρουσιαστούν σε δισδιάστατο χώρο, προτείνει να παρουσιάζονται πάντα φέτες από δεδομένα του k-διάστατου χώρου. Το συγκεκριμένο τμήμα δεδομένων που θα προβάλλεται κάθε φορά προσδιορίζεται από το χρήστη αρχικά αλλά και δυναμικά μέσα από το γράφημα μπορεί να γίνει αλλαγή του σημείου αυτού.

Η τεχνική στηρίζεται στην ιδέα ότι ο χρήστης θέλει να δει κάθε φορά ένα συγκεκριμένο σημείο και μια περιοχή γύρω από αυτό. Ορίζει λοιπόν, ένα σημείο c το οποίο είναι το επίκεντρο του ενδιαφέροντος του χρήστη και ένα εύρος τιμών (ακτίνα) w. Υποθέτοντας ότι θέλουμε να αναπαραστήσουμε δεδομένα k διαστάσεων, τότε το σημείο c ορίζεται ως $c=f(c_1, c_2, c_3, \dots, c_k)$. Στον πίνακα των Scatterplots παρουσιάζουμε τη συνάρτηση f διατηρώντας σταθερές όλες τις μεταβλητές που προσδιορίζουν τη συνάρτηση εκτός από τις δύο μεταβλητές που παρουσιάζονται, π.χ. έστω ότι παρουσιάζονται οι μεταβλητές (x,y) τότε το γράφημα θα σχεδιαστεί από τη συνάρτηση $f(x_i, y_i, c_3, c_4, \dots, c_k)$, όπου $c_i, i=3..k$, είναι σταθερές τιμές που προσδιορίζουν το σημείο c και x_i, y_i παίρνουν ως τιμές όλο το πεδίο ορισμού τους (Εικόνα 8,9).

Μειονεκτήματα: εξακολουθεί να ισχύει η κατάχρηση του διαθέσιμου χώρου στην οθόνη.



Εικόνα 8



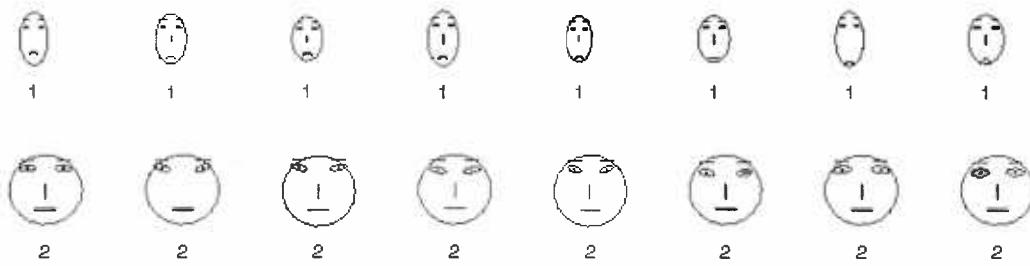
Εικόνα 9

2.3.3 Τεχνικές Βασιζόμενες σε Εικόνες

Η βασική ιδέα αυτών των τεχνικών είναι η παρουσίαση των αποτελεσμάτων χρησιμοποιώντας εικονίδια. Κάθε μονάδα των δεδομένων αντιπροσωπεύεται από κάποιο εικονίδιο, το οποίο έχει επιλεγεί με βάση κάποιους συγκεκριμένους κανόνες ώστε να εκφράζει τα χαρακτηριστικά της κάθε μονάδας. Υπάρχουν διάφορες παραλλαγές των τεχνικών αυτών, οι οποίες περιγράφονται ακολούθως:

Πρόσωπα Chernoff

Η τεχνική αυτή είναι αρκετά παλιά και γνωστή. Παρουσιάζει τα αποτελέσματα της σε δισδιάστατο χώρο και τις υπόλοιπες διαστάσεις τις αντιστοιχεί σε διαφορετικές μορφές προσώπων [33] (Εικόνα 10), π.χ. το σχήμα του προσώπου, των ματιών, της μύτης κ.λ.π.



Εικόνα 10

Πλεονεκτήματα: Στηρίζεται στο γεγονός ότι ο άνθρωπος είναι εξοικειωμένος με τις εκφράσεις του προσώπου, συνεπώς είναι εύκολα κατανοητό το αποτέλεσμα.

Μειονεκτήματα: το μέγεθος των δεδομένων που μπορούν να αναπαρασταθούν είναι περιορισμένο. Μεγάλος όγκος δεδομένων μπορεί να οδηγήσει σε επικάλυψη των εικονιδίων με αποτέλεσμα να χαθούν οι μορφές των προσώπων.

Ένα ακόμη μειονέκτημα της τεχνικής αυτής είναι ότι δεν παρουσιάζει πληροφορία για την πραγματική τιμή των δεδομένων. Το σχήμα του προσώπου βοηθάει στο να χαρακτηρίσεις τα δεδομένα με σχετική ευκολία, αλλά όχι στο να λάβεις λεπτομερή ανάλυση των τιμών τους.

Ο πίνακας που ακολουθεί δηλώνει ένα προτεινόμενο τρόπο παρουσίασης των διαστάσεων σε σχέση με τα χαρακτηριστικά των προσώπων:

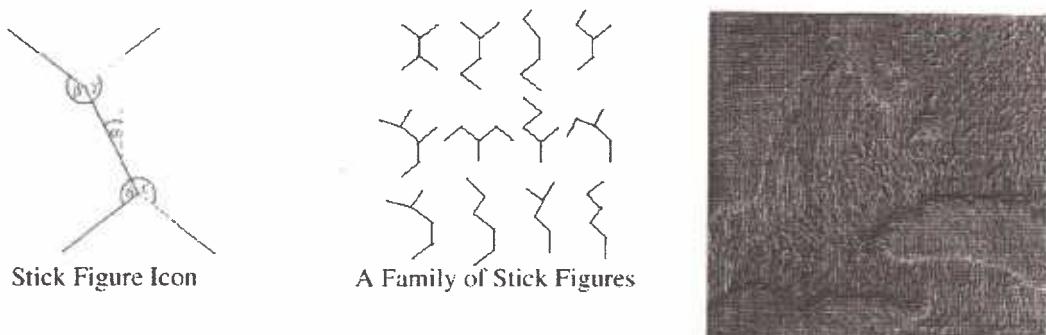
Διάξταση	Ιδιότητα προσώπου
1	Face width
2	Ear level
3	Half face height
4	Eccentricity of upper ellipse of face
5	Eccentricity of lower ellipse of face
6	Length of nose
7	Position of center of mouth
8	Curvature of mouth
9	Length of mouth
10	Height of center of eyes
11	Separation of eyes
12	Slant of eyes
13	Eccentricity of eyes
14	Half length of eye
15	Position of pupil
16	Height of eyebrow
17	Angle of brow
18	Length of brow
19	Radius of ear
20	Nose width

Γραμμικά Σχήματα (Stick Figures)

Η τεχνική αυτή μοιάζει με την τεχνική των Chernoff faces, αλλά υπερτερεί στον όγκο των δεδομένων που μπορεί να αναπαραστήσει, ώστε να είναι πιο κατάλληλη για τις ανάγκες του Data Mining.

Στην περίπτωση των stick figures δεν χρησιμοποιούνται πρόσωπα αλλά μικρές σε όγκο πολυγωνικές γραμμές οι οποίες εμπεριέχουν πληροφορία για τα

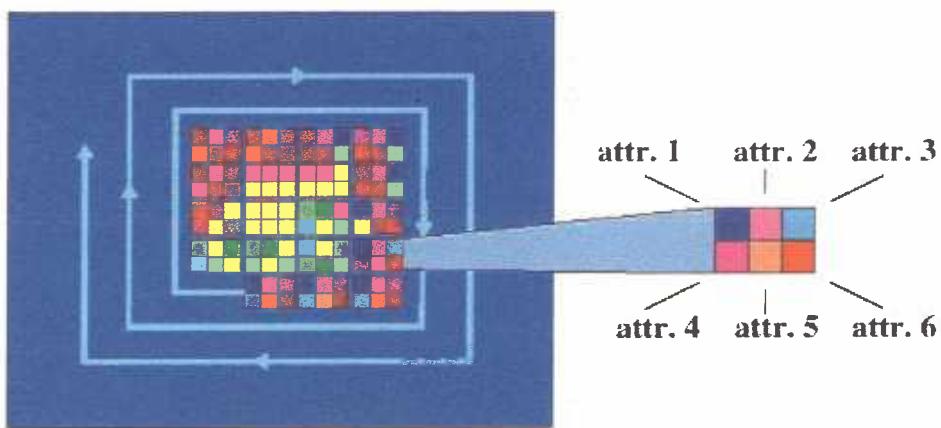
χαρακτηριστικά των δεδομένων που παρουσιάζουν μέσα στο σχήμα τους [32]. Συγκεκριμένα, χρησιμοποιείται ένα δισδιάστατο γράφημα για την απεικόνιση των δύο διαστάσεων, ενώ οι υπόλοιπες διαστάσεις απεικονίζονται στα χαρακτηριστικά της γραμμής, όπως: γωνίες, μήκος τμημάτων της πολυγωνικής γραμμής, πάχος τμημάτων γραμμής κ.λ.π. (Εικόνα 11). Η τοποθέτηση τέτοιων γραμμών στο δισδιάστατο χώρο δημιουργεί πρότυπα-υποδείγματα (patterns) τα οποία μας δίνουν χρήσιμη πληροφόρηση π.χ. όταν δούμε ότι όλο το γράφημα έχει μια ομοιόμορφη κατανομή - ομοιόμορφο πρότυπο, αυτό σημαίνει ότι στην k διάσταση (όπου $k > 2$) υπάρχει μια ομοιογένεια στα δεδομένα μας.



Εικόνα 11

Μορφοποίηση Σχήματος (Shape Coding)

Τα δεδομένα παρουσιάζονται σε μια διάσταση και οι υπόλοιπες διαστάσεις παρουσιάζονται σε μικρούς πίνακες από πεδία όπου το κάθε πεδίο αποτελεί μία διάσταση. Ως παράδειγμα, αν έχουμε ένα σύνολο δεδομένων με k διαστάσεις, τότε μια διάσταση θα αποτελεί την διάσταση με βάση την οποία τοποθετούνται οι πίνακες στην οθόνη και οι $k-1$ διαστάσεις θα αποτελούν πεδία των πινάκων (Εικόνα 12). Ο τρόπος τοποθέτησης των πινάκων στην οθόνη είναι απλός και ακολουθεί τη λογική γεμίσματος της οθόνης ανά γραμμή δηλ, τοποθετώντας τους πίνακες από αριστερά



Εικόνα 12

προς τα δεξιά και στην επόμενη γραμμή από δεξιά προς τα αριστερά κ.ο.κ. Θα μπορούσε να χρησιμοποιηθεί και η τεχνική γεμίσματος ανά στήλη. Η τεχνική αυτή αποτελεί το προκαρυωτικό στάδιο των Pixel oriented τεχνικών.

Έγχρωμη Μορφοποίηση Σχήματος (Color Icons)

Η τεχνική αυτή είναι ίδια με την τεχνική shape coding με την διαφορά ότι χρησιμοποιούνται χώματα για την απεικόνιση των τιμών των δεδομένων, σε αντίθεση με την shape coding που χρησιμοποιεί αποχρώσεις του γκρίζου (Εικόνα 12).

Τεχνική Αναζήτησης Κειμένων – TileBars (Πλακάκια)

Τα TileBars χρησιμοποιούνται κυρίως για αναζήτηση κειμένων (text retrieval) και σκοπός τους είναι να δείξουν στον χρήστη την ομοιότητα που υπάρχει ανάμεσα στην ερώτηση που έθεσαν και τα κείμενα που επιστράφηκαν ως αποτέλεσμα [14][15][16][17]. Η ερώτηση που τίθεται έχει τη μορφή λέξεων οι οποίες χαρακτηρίζουν αυτό που ψάχνει ο χρήστης, π.χ. οι λέξεις ιατρική, νοσοκομείο, περιθαλψη κάνουν σαφές το αντικείμενο στο οποίο αναφέρεται ο χρήστης. Όσο πιο συναφείς είναι οι λέξεις που χρησιμοποιούνται τόσο πιο σχετικό θα είναι το αποτέλεσμα της ερώτησης δηλ. τόσο πιο σχετικά θα είναι μεταξύ τους αλλά και ως προς την ερώτηση τα κείμενα που θα βρεθούν.

Το γραφικό αποτέλεσμα βοηθάει το χρήστη να καταλάβει ποια κείμενα περιέχουν και σε τι βαθμό αυτό που ζητήθηκε. Οι στόχοι της τεχνικής είναι:

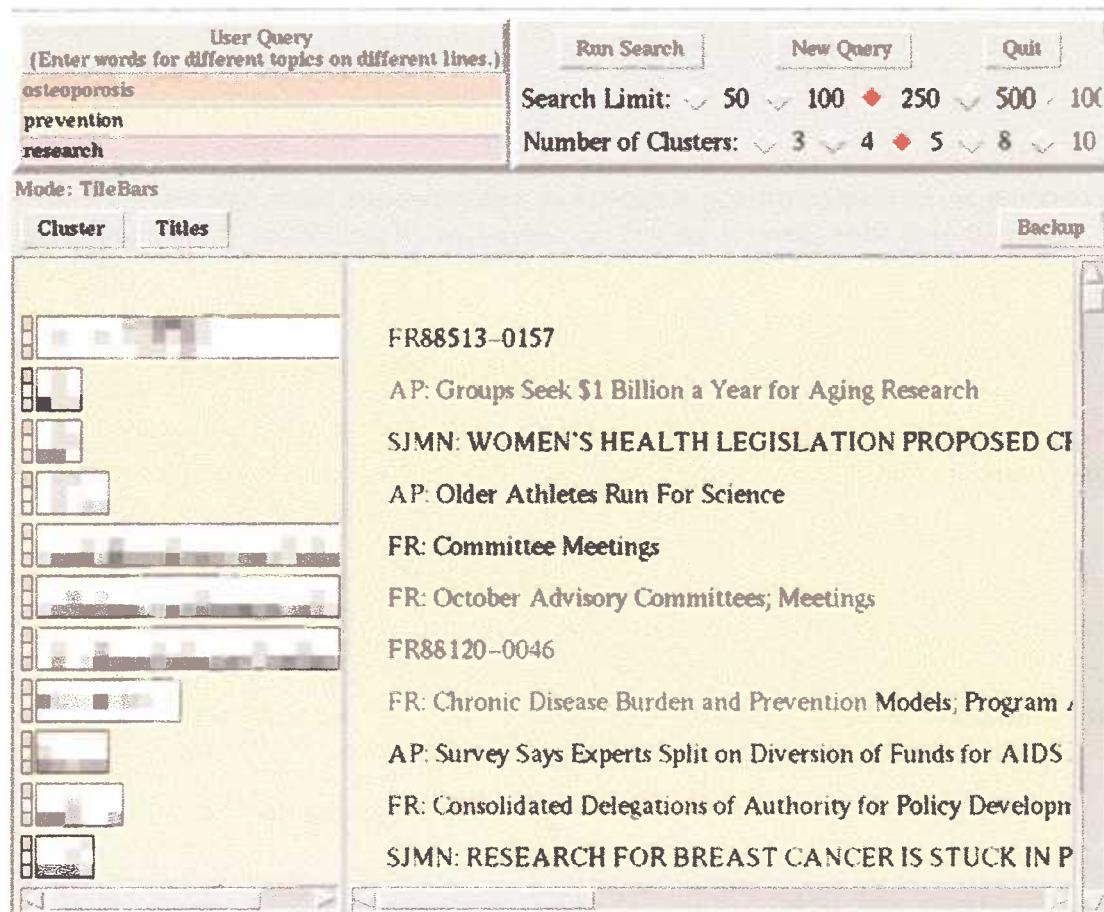
- Ο βαθμός συσχέτισης του κάθε επιλεγμένου κειμένου,
- Η συχνότητα εμφάνισης των λέξεων που αναζητήθηκαν μέσα στο συγκεκριμένο κείμενο,
- Την κατανομή των λέξεων που ζητήθηκαν μέσα στο κείμενο σε σχέση με τα άλλα επιλεγόμενα κείμενα.

Κατά τη διαδικασία εύρεσης των λέξεων, το κάθε κείμενο χωρίζεται σε τμήματα μέσα στα οποία γίνεται η αναζήτηση της κάθε λέξης που ζητήθηκε από τον χρήστη. Έτσι, για κάθε έγγραφο που επιλέγεται, η γραφική απεικόνιση του αποτελέσματος είναι ένα παραλληλόγραμμο (Εικόνα 13) το οποίο είναι χωρισμένο σε



Εικόνα 13

τόσες γραμμές όσες είναι και οι λέξεις που ζητήθηκαν και σε τόσες στήλες όσα είναι τα τμήματα στα οποία χωρίστηκε το κάθε έγγραφο. Ανάλογα με το βαθμό συσχέτισης της κάθε λέξης με το κάθε τμήμα του κειμένου, χρωματίζεται το αντίστοιχο τετραγωνάκι του παραλληλογράμμου όπως διαφαίνεται στο παρακάτω παράδειγμα (Εικόνα 14).



Εικόνα 14

Στο παραπάνω παράδειγμα έχουν ζητηθεί οι λέξεις "osteoporosis, prevention, research". Το αποτέλεσμα περιλαμβάνει τα κείμενα που βρέθηκαν, και για κάθε κείμενο υπάρχει ένα Tile Bar. Το κάθε Tile Bar αποτέλείται από 3 γραμμές (μία για κάθε λέξη που ζητήθηκε) και από πολλές στήλες, οι οποίες δηλώνουν τα διαφορετικά τμήματα στα οποία χωρίστηκε το έγγραφο. Οι αποχρώσεις του γκρίζου δείχνουν το βαθμό συσχέτισης της κάθε λέξης με το κάθε τμήμα. Το μαύρο χρώμα σημαίνει μεγάλη συσχέτιση.

2.3.4 Ιεραρχικές Τεχνικές

Οι τεχνικές αυτές παρουσιάζουν τα δεδομένα χρησιμοποιώντας μια ιεραρχική κατηγοριοποίηση της οθόνης σε υπό-τμήματα. Τέτοιες τεχνικές είναι οι ακόλουθες:

Ιεραρχική Κατανομή (Dimensional Stacking)

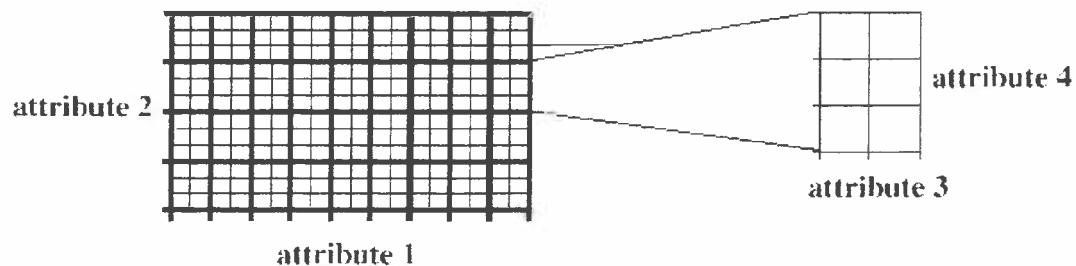
Η τεχνική αυτή [12][13] παρουσιάζει τις k-διαστάσεις των δεδομένων στον δισδιάστατο χώρο. Τεμαχίζει τον διαθέσιμο χώρο έτσι ώστε τα κατώτερα επίπεδα να συνθέτουν το αμέσως ανώτερο επίπεδο. Δηλ. αν έχουμε

- X επίπεδα ιεραρχίας και
- N μεταβλητές,

τότε αν τα γνωρίσματα N1, N2 ορίζουν το κατώτερο επίπεδο ιεραρχίας, τα γνωρίσματα N3, N4 ορίζονται από κοινού από τα γνωρίσματα N1, N2 και

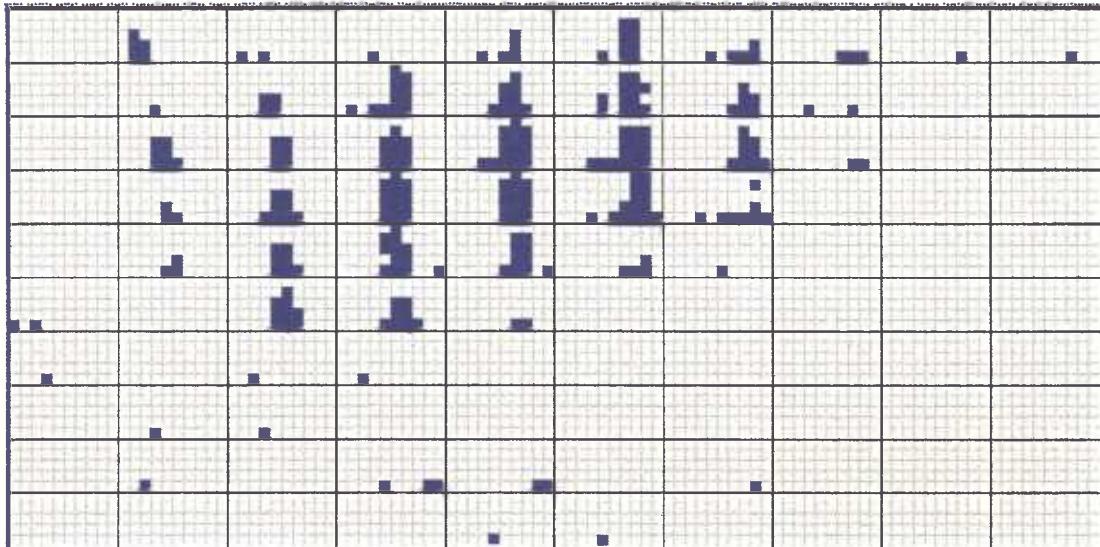
- N1, N2, N3, N4 = N και
- X=2.

Οι τιμές των δεδομένων μπορούν να αναπαρασταθούν με χρώματα. Η συγκεκριμένη τεχνική απευθύνεται κυρίως σε δεδομένα με ιεραρχική διάταξη π.χ. αν ένα γνώριμα N2 περιγράφει τους νομούς της Ελλάδας και κάποιο άλλο γνώριμα N1 περιγράφει τις επαρχίες των νομών τότε το N2 σχηματίζεται από ένα σύνολο τιμών του γνωρίσματος N1. Το ακόλουθο σχήμα περιγράφει την παραπάνω λογική. Να σημειωθεί ότι τα γνωρίσματα 3-4 είναι ιεραρχικά κατώτερα από τα γνωρίσματα 1-2.



Εικόνα 15

Παράδειγμα: Στη Εικόνα 16 παρουσιάζονται πληροφορίες εξόρυξης πετρελαιοειδών, όπου στους εξωτερικούς άξονες x-y παρουσιάζεται το γεωγραφικό μήκος και πλάτος και στους εσωτερικούς άξονες παρουσιάζεται η ποιότητα του ορυκτού και βάθος εξόρυξης του.



Εικόνα 16

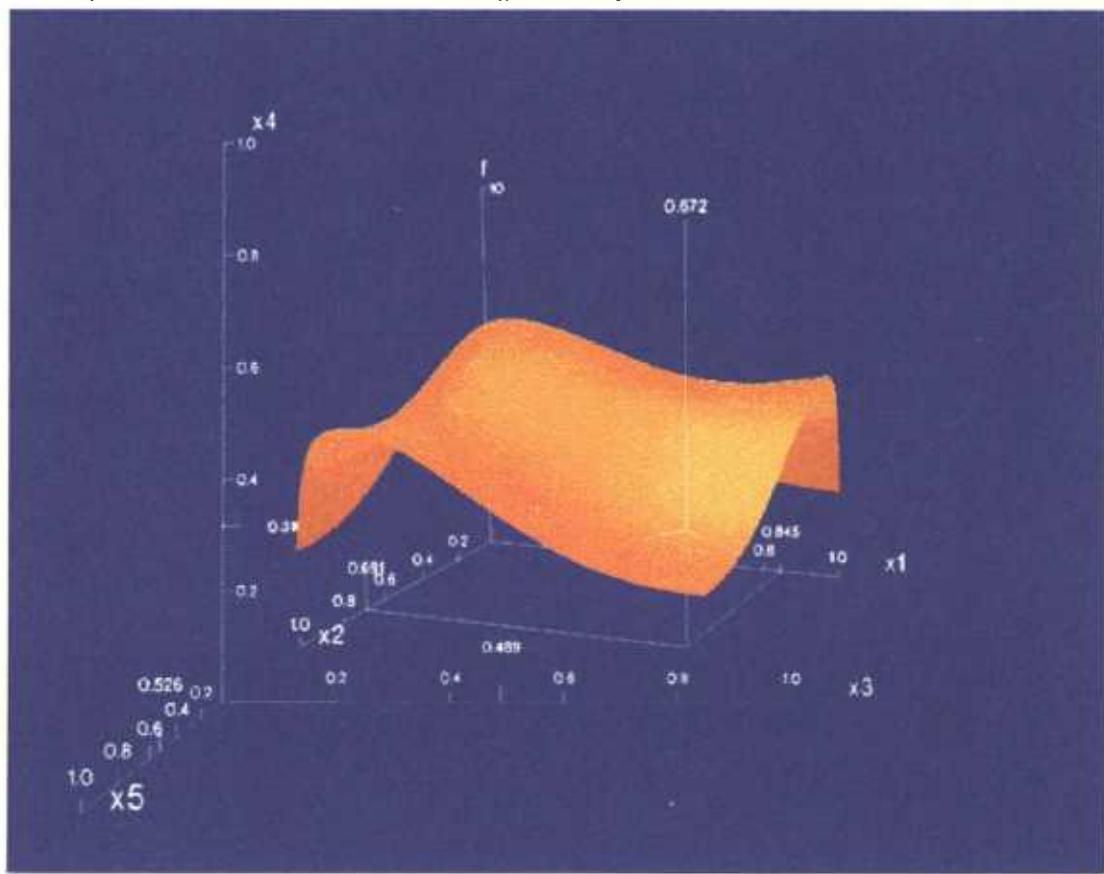
Μειονεκτήματα: Απαιτείται ζυγός αριθμός γνωρισμάτων, δεδομένου ότι η παρουσίαση πραγματοποιείται σε δισδιάστατο σύστημα αξόνων.

Ενθυλακωμένα Διαγράμματα – Worlds within Worlds

Η τεχνική αυτή παρουσιάζει τα δεδομένα σε τρισδιάστατο χώρο και είναι επίσης γνωστή ως n-Vision τεχνική [9][10][11]. Σε κάθε μία από τις τρεις διαστάσεις παρουσιάζονται περισσότερα από ένα γνωρίσμα. Η μέθοδος αυτή, προκειμένου να δημιουργήσει όσο το δυνατόν πιο κατανοητό αποτέλεσμα, για να αναπαραστήσει ένα σύνολο από γνωρίσματα τα οποία δεν έχουν την ίδια σύνθεση και λογική (multivariate), παράγει δυναμικά το γράφημα.

Συγκεκριμένα, ένας τρόπος για να μειωθεί η πολυπλοκότητα των multivariate συναρτήσεων είναι να τεθούν μερικές από τις ανεξάρτητες (independent) μεταβλητές σταθερές για μια συγκεκριμένη τιμή τους. Κάθε σταθερή τιμή κάθε ανεξάρτητης μεταβλητής αντιστοιχεί σε μία μικρή φέτα του πολυδιάστατου χώρου για αυτή την μεταβλητή. Κάνοντας το ίδιο και για τις k-2 μεταβλητές, όπου k το σύνολο των μεταβλητών, καταλήγουμε σε ένα τρισδιάστατο γράφημα (2 μεταβλητές και στην τρίτη διάσταση το μέγεθος που παρακολουθείται) το οποίο μπορεί εύκολα να αναπαρασταθεί με γνωστές τεχνικές απεικόνισης γραφημάτων. Τις υπόλοιπες μεταβλητές που θέσαμε να είναι σταθερές τις απεικονίζουμε με ένα μεγαλύτερο σύστημα τριών αξόνων, μέσα στο οποίο βρίσκεται φωλιασμένο το ήδη σχεδιασμένο γράφημα. Στην ουσία η τεχνική αυτή υποχρεώνει να επιλέξουμε ένα σημείο στον πολυδιάστατο χώρο και στη συνέχεια να παρατηρήσουμε τη συμπεριφορά 2 μεταβλητών σε σχέση με αυτό. Επιλέγοντας ένα νέο σημείο στο εξωτερικό σύστημα αξόνων αλλάζουν οι τιμές των σταθερών γνωρισμάτων και επανασχεδιάζεται το εσωτερικό γράφημα. Η λογική αυτή μπορεί να εφαρμοστεί ακολουθιακά μέχρι να απεικονιστούν όλα τα γνωρίσματα. Στην ουσία δημιουργείται ένα interactive master-detail σύστημα.

Ένα παράδειγμα είναι το ακόλουθο: έστω η συνάρτηση $f(x_1, x_2, x_3, x_4, x_5)$ όπου τα x_1-5 είναι γνωρίσματα της βάσης και η συνάρτηση f παράγει τις τιμές τους (στην ουσία η f θα μπορούσε να θεωρηθεί ως μία μεταβλητή με τιμές που επιθυμούμε να εξετάσουμε όπως π.χ. πωλήσεις). Θέτουμε σταθερά τα x_3, x_4, x_5 με τις ακόλουθες τιμές αντίστοιχα c_3, c_4, c_5 . Τώρα πρέπει να αναπαραστήσουμε τη συνάρτηση $f(x_1, x_2, c_3, c_4, c_5)$, θέτοντας τις τιμές του x_1 στον x_1 άξονα, τις τιμές του x_2 στον x_2 άξονα και τις τιμές της συνάρτησης στον z άξονα. Οι υπόλοιπες τρεις τιμές θα αναπαρασταθούν σε ένα νέο σύστημα 3 αξόνων το οποίο θα ενθυλακώνει το



Εικόνα 17

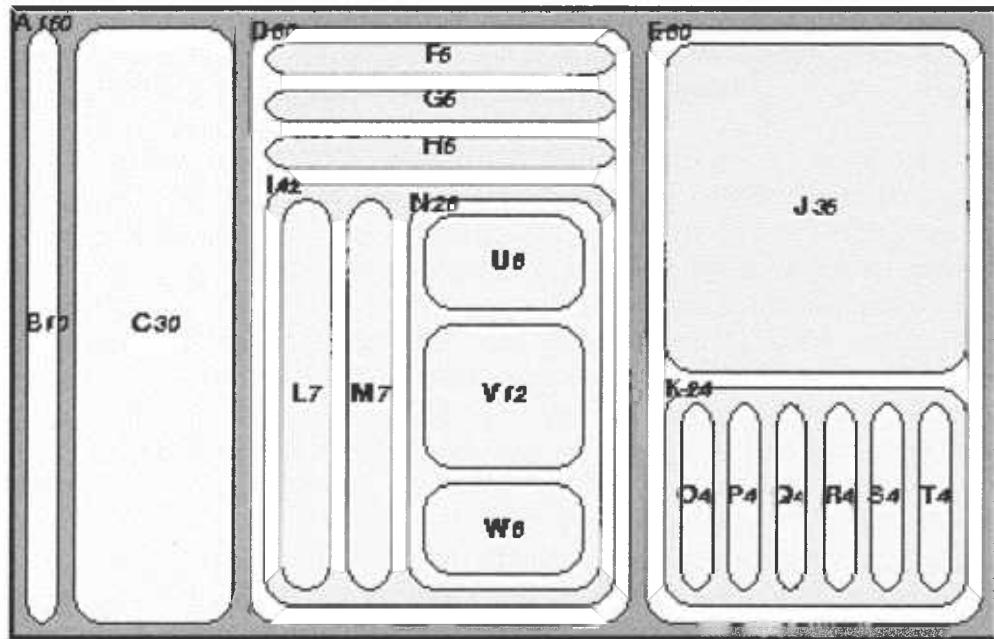
μικρότερο ήδη δημιουργημένο γράφημα (Εικόνα 17). Οι τιμές που έχουν επιλεγεί για τα γνωρίσματα x_3 , x_4 , x_5 εμφανίζονται στο εξωτερικό σύστημα αξόνων με μεγαλύτερη γραμματοσειρά. Δυναμικά μπορεί να επιλεγεί νέο σύνολο τιμών για τα σταθερά γνωρίσματα και να ζητηθεί η επανασχεδίαση του εσωτερικού γραφήματος. Οι επιλογή των σταθερών και των μεταβλητών γνωρισμάτων γίνεται από το χρήστη και μπορεί να αλλάξει δυναμικά.

Πλεονεκτήματα: Αποτελεί μία εύκολα κατανοητή τεχνική και παρέχει τη δυνατότητα στο χρήστη να αλλάξει δυναμικά τα δεδομένα.

Μειονεκτήματα: Γίνεται πολύπλοκη για την ανθρώπινη σκέψη όταν αυξάνονται οι διαστάσεις των δεδομένων. Επίσης, δεν παρέχει τη δυνατότητα παρακολούθησης όλων των τιμών των δεδομένων ταυτόχρονα αφού πάντα παρουσιάζει μία φέτα αυτών στον πολυδιάστατο χώρο. Ως αποτέλεσμα είναι, ο χρήστης να μην έχει μια συνολική - συγκεντρωτική εικόνα για τα δεδομένα του στο ανώτερο ιεραρχικό επίπεδο.

Δεντρική - Ενθυλακωμένη Παράθεση (Treemap – Venn)

Η τεχνική αυτή [7][8] είναι σχεδιασμένη ώστε να παρουσιάζει δεδομένα που εμπεριέχουν μια δομημένη ιεραρχικά πληροφορία. Δημιουργείται τεμαχίζοντας την οθόνη σε ενθυλακωμένα παραλληλόγραμμα τα οποία παρουσιάζουν την ιεραρχική δομή των δεδομένων. Τα ίδιαίτερα χαρακτηριστικά του κάθε παραλληλόγραμμου παρουσιάζονται από το σχήμα του και το χρώμα γεμίσματός του. Οι Treemap τεχνικές χρησιμοποιούνται για να παρουσιάσουν σχετική πληροφόρηση των δομών των δεδομένων μέσα στις ιεραρχίες.



Εικόνα 18

Γενικές ιδιότητες των Treemap διαγραμμάτων:

- Κάθε κόμβος παιδί περιλαμβάνεται ολοκληρωτικά ή είναι ίσος με τον κόμβο πατέρα
- Δύο παραλληλόγραμμα δύο κόμβων τέμνονται μόνο όταν ο ένας κόμβος είναι παιδί του άλλου
- Κάθε κόμβος καταλαμβάνει μια περιοχή σε αυστηρή σχέση με το βάρος του
- Το βάρος του κάθε κόμβου είναι μεγαλύτερο ή ίσο του αθροίσματος των βαρών των παιδιών του

Ο όρος βάρος, που χρησιμοποιήθηκε παραπάνω ορίζει το μέγεθος του κάθε αντικειμένου και είναι υποχρεωτικός ο ορισμός του. Το βάρος είναι ένα μετρούμενο μέγεθος των δεδομένων π.χ θα μπορούσε να σημαίνει "αριθμός ετών", που ένας υπάλληλος δουλεύει σε κάποια εταιρεία.

Επίσης, η τεχνική αυτή χρησιμοποιεί και το χρώμα γεμίσματος για να παρουσιάσει κάποιο άλλο μετρούμενο μέγεθος των δεδομένων. Στο προηγούμενο παράδειγμα θα μπορούσε να δείχνει το μέγεθος του μισθού (Εικόνα 18).

Υπάρχουν δύο οικογένειες αλγορίθμων, οι οποίες προσδιορίζουν τον τρόπο με τον οποίο γίνεται ο καταμερισμός χώρου στην οθόνη:

- **Slice and dice:** Ο αλγόριθμος αυτός τεμαχίζει την οθόνη σε παραλληλόγραμμα, τα οποία αναπαριστούν αντικείμενα του δέντρου, επαναλαμβανόμενα, αλλάζοντας την κατεύθυνση τεμαχισμού με κάθε αλλαγή επιπέδου.

Αν θεωρηθεί ότι το iερταρχικό δέντρο είναι μια γεωγραφική κατανομή των πωλήσεων ανά γεωγραφικά διαμερίσματα της Ελλάδας τότε θα έχουμε την ακόλουθη κατάτυπη της οθόνης: θεωρούμε ότι το σύνολο της οθόνης αναπαριστά την Ελλάδα και δίνουμε ένα χρώμα ανάλογα με το σύνολο των πωλήσεων, στη συνέχεια χωρίζουμε την Οθόνη σε ν νομούς όπου ως βάρος θέτουμε το γεωγραφικό μέγεθος του κάθε νομού. Ο κάθε νομός είναι ένα παραλληλόγραμμο το οποίο ξεκινάει από την κορυφή της οθόνης μέχρι το κάτω μέρος της, δηλ. καταλαμβάνει όλο το ύψος του κόμβου πατέρα αλλά το πλάτος του ισοδυναμεί με το ποσοστό που του αντιστοιχεί με βάση το βάρος που έχει τεθεί (γεωγραφικό μέγεθος του κάθε νομού). Στο επόμενο επίπεδο, το οποίο καταλαμβάνουν οι επαρχίες, κάθε παραλληλόγραμμο - νομός τεμαχίζεται οριζοντιώς, δηλ. η κάθε επαρχία καταλαμβάνει όλο το εύρος του νομού που ανήκει ως προς το πλάτος αλλά μόνο το ποσοστό που της αντιστοιχεί ως προς το ύψος. Η διαδικασία αυτή εκτελείται επαναληπτικά (Εικόνα 18).

Υπάρχουν δύο υπό-κατηγορίες αλγορίθμων:

- **Nested:** Το σύνολο των κόμβων παιδιά αφήνει μία μικρή περιοχή περιμετρικά (offset) η οποία παρουσιάζει τον πατέρα κόμβο και
- **Not Nested:** όπου συμβαίνει το αντίθετο, δηλ. τα παιδιά δεν αφήνουν καθόλου περιθώρια από τον κόμβο πατέρα, με αποτέλεσμα το διάγραμμα να παρουσιάζει μόνο τους τερματικούς κόμβους.

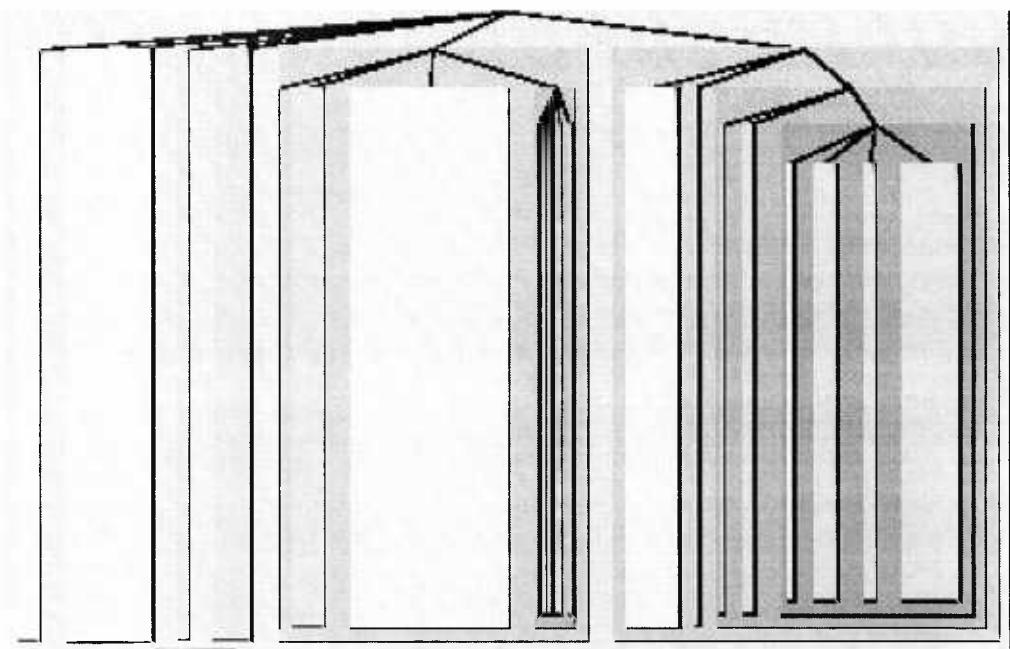
- **Top Down:** Οι αλγόριθμοι αυτοί είχαν σχεδιαστεί με βάση την ιδέα των δέντρων και ακολουθούν παρόμοιο τρόπο παρουσίασης των αποτελεσμάτων (Εικόνα 19). Ο αλγόριθμος TOP-DOWN περιγράφεται ακολούθως:

- i. Θεώρησε το σύνολο της οθόνης ως τον κόμβο «ρίζα»
- ii. Χώρισε τον χώρο του κόμβου σε ν κάθετα τμήματα, όπου ν το σύνολο των παιδιών, ανάλογα με το βάρος του κάθε παιδιού (ποσοστό κάλυψης επί του συνόλου). Δηλ. υπολόγισε το σύνολο των βαρών όλων των



παιδιών και στη συνέχεια υπολόγισε το ποσοστό του κόμβου που αντιστοιχεί σε κάθε παιδί.

- iii. Όλα τα παιδιά του κόμβου σχεδιάζονται στον οριζόντιο άξονα ένα επίπεδο πιο κάτω από αυτό του πατέρα.
- iv. Σχεδίασε μια περιοχή μέσα στον κόμβο πατέρα που να αντιστοιχεί σε ένα παιδί και κάνε το ίδιο για όλα τα παιδιά του κόμβου.
- v. Θέσε σαν ενεργό ένα παιδί κάθε φορά και πήγαινε στο βήμα ii για όλα τα παιδιά του κόμβου.



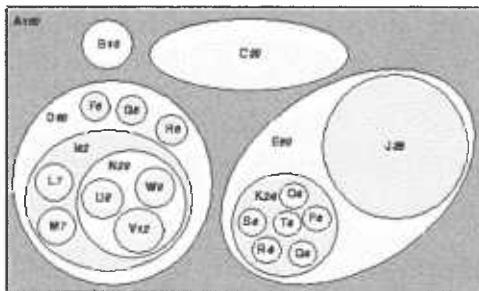
Top-Down, Size by Weight

Εικόνα 19

Μειονεκτήματα: ο αλγόριθμος Top-Down είναι ότι προκειμένου να παρουσιάσει δεδομένα με μεγάλες ιεραρχίες, απαιτεί μεγάλο “ύψος” στην οθόνη, με αποτέλεσμα τα δεδομένα είτε να είναι δυσανάγνωστα είτε να μην αναπαρίστανται στο σύνολό τους. Το πρόβλημα αυτό επιλύεται, εν μέρει, από τον αλγόριθμο slice and dice, οποίος μπορεί να αναπαραστήσει μέχρι και 1000 αντικείμενα στην οθόνη.

Πλεονεκτήματα: ο αλγόριθμος Top-Down παρουσιάζει μεγάλη οπτική συνάφεια με την τεχνική των δέντρων, έχοντας ως άμεσο αποτέλεσμα την εύκολη κατανόησή του από τον χρήστη.

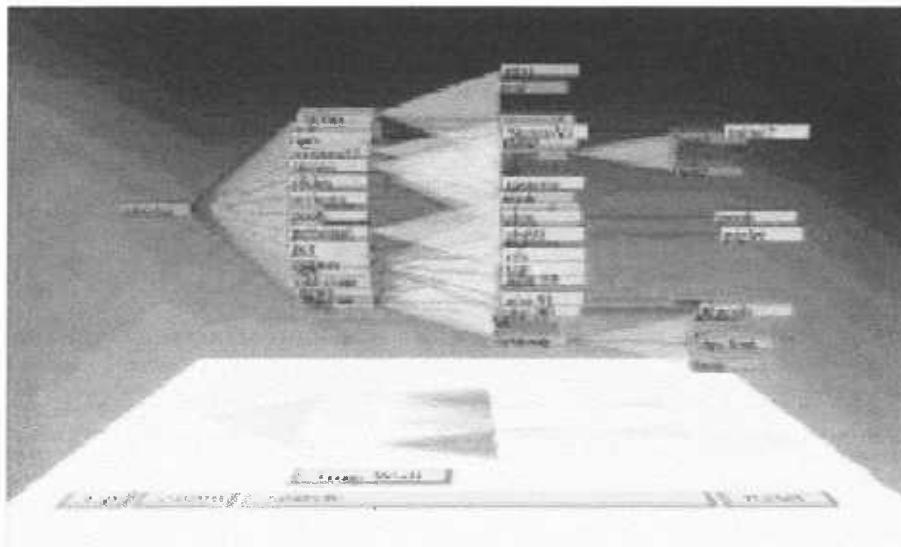
Παρατήρηση: Η τεχνική Treemap έχει λογική συνάφεια με τα Venn Διαγράμματα (Εικόνα 20).



Εικόνα 20

Κωνικά Δέντρα (Cone Trees)

Η τεχνική των Cone Trees [3] είναι μια εξέλιξη των απλών δέντρων και αποσκοπεί στην καλύτερη εκμετάλλευση χώρου για αναπαράσταση ιεραρχικών δομών δεδομένων. Χρησιμοποιεί τεχνολογία τρισδιάστατων γραφικών, όπου κάθε αντικείμενο-κόμβος παρουσιάζεται με κείμενο σε ένα τετράγωνο στην οθόνη και όλα τα παιδιά του βρίσκονται κάτω από αυτό σε σχηματισμό κώνου (Εικόνα 21). Η ρίζα



Εικόνα 21

(πατέρας όλων) τοποθετείται στην κορυφή της οθόνης και το επόμενο επίπεδο τοποθετείται πιο κάτω με τρόπο ώστε:

- Το τελευταίο επίπεδο του δέντρου να βρίσκεται κοντά στο κάτω μέρος της οθόνης (Χρήση της οθόνης στο μέγιστο βαθμό).
- Το aspect ratio του δέντρου προσδιορίζεται έτσι ώστε όλο το δέντρο να παρουσιάζεται στο σύνολο της οθόνη.

- Κάθε κώνος του δέντρου έχει το ίδιο ύψος (το οποίο είναι ίσο με το "ύψος της οθόνης / το σύνολο των επιπέδων του δέντρου").
- Τα αντικείμενα-κόμβοι τα οποία βρίσκονται στο πίσω μέρος του κώνου παρουσιάζονται μόνο αν δεν καλύπτονται από κάποιο κώνο στο εμπρός μέρος του κώνου.
- Αν η περιγραφή κάποιου κόμβου δεν χωράει στο τετράγωνο που έχει οριστεί τότε αυτή παρουσιάζεται μόνο όταν επιλεγεί.
- Όταν επιλεγεί κάποιος κόμβος φύλλο, τότε τονίζεται όλο το μονοπάτι μέχρι τη ρίζα. Αν σε κάποιο επίπεδο ο κόμβος που αντιστοιχεί στο επιλεγμένο μονοπάτι βρίσκεται στο πίσω μέρος του κώνου τότε ο κώνος περιστρέφεται μέχρις ότου να έρθει στο προσκήνιο ο συγκεκριμένος κόμβος.
- Ο κάθε κώνος μπορεί να περιστρέφεται συνεχώς μέχρι ο χρήστης να βρει τον κόμβο που τον ενδιαφέρει.

Πλεονεκτήματα: Η εισαγωγή της τρισδιάστατης τεχνολογίας στην τεχνική των δέντρων βελτιώνει το αποτέλεσμα κατά μεγάλο βαθμό. Συγκεκριμένα, αν υποτεθεί ότι έχουμε ένα δέντρο με L επίπεδα και ο branching factor είναι b τότε το πλάτος της βάσης θα είναι b^{L-1} και το aspect ratio θα είναι b^{L-1}/L . Αυτό σημαίνει ότι το κλασικό δέντρο αυξάνει το μέγεθός του εκθετικά όσο μεγαλώνει ο branching factor. Αντίθετα, το cone tree προποιούντας το πάχος του κώνου μπορεί να αναπαριστά εύκολα μεγάλο όγκο δεδομένων.

Τέλος, η τεχνική αυτή χρησιμοποιεί και την λειτουργικότητα της fisheye τεχνικής χωρίς να χρειαστεί να προσδιορίσει το βαθμό ενδιαφέροντος χρησιμοποιώντας αντίστοιχες συναρτήσεις. Απλά υπογραμμίζει την εκάστοτε επιλογή.

Iεραρχικός Κύβος (InfoCube)

Η τεχνική αυτή [6] χρησιμοποιεί τρισδιάστατα γραφικά για την αναπαράσταση της ιεραρχικής πληροφορίας και επιτρέπει την αλληλεπίδραση με το χρήστη.

Αποτελείται από ενθυλακωμένους κύβους οι οποίοι αναπαριστούν την ιεραρχική δομή και ο κάθε κύβος αναπαριστά ένα κόμβο-αντικείμενο (Εικόνα 22). Ο ανώτερος ιεραρχικά κύβος (και συνεπώς ο πιο εξωτερικός) αναπαριστά τα δεδομένα στο σύνολό τους.

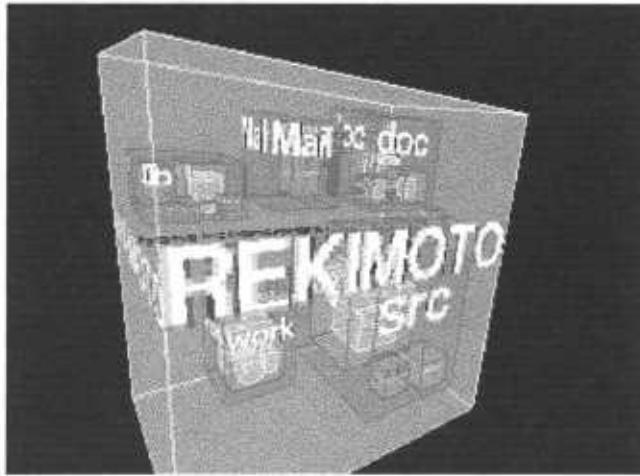
Οι κύβοι είναι ημιδιαφανείς, έτσι ώστε να έχει τη δυνατότητα ο χρήστης να δει τους εσωτερικούς κύβους. Πάνω σε κάθε κύβο υπάρχει ένα λεκτικό το οποίο περιγράφει το αντικείμενο που παρουσιάζεται.

Ο κύβος μπορεί να περιστραφεί για καλύτερη παρουσίαση και διερεύνηση. Τέλος παρέχεται η δυνατότητα drill down προκειμένου να μπορέσει ο χρήστης να εξετάσει με μεγαλύτερη λεπτομέρεια τα δεδομένα του κατώτερου επιπέδου.



Το μέγεθος και το χρώμα του κύβου μπορούν να χρησιμοποιηθούν για την παρουσίαση κάποιου μετρούμενου-εξεταζόμενου μεγέθους.

Πλεονεκτήματα: η τεχνική αυτή μπορεί εύκολα να αναπαραστήσει μεγάλο αριθμό κόμβων παιδιών για κάθε κόμβο πατέρα δεδομένου ότι είναι υλοποιημένη σε πραγματικό τρισδιάστατο περιβάλλον. Ετσι, αν υποθέσουμε ότι έχουμε 1000 κόμβους, πρέπει να αναπαρασταθούν 10 αντικείμενα σε κάθε διάσταση δηλ. $10 \times 10 \times 10 = 1000$. Επίσης, δεν υπάρχει όριο όσον αφορά τον αριθμό επιπέδων της ιεραρχίας γιατί δίνεται δυνατότητα drill down στον κύβο, συνεπώς τα κατώτερα επίπεδα δεν είναι αναγκαίο να αναπαρασταθούν από την αρχή, αλλά μπορούν να σχηματίζονται σταδιακά σε κάθε drill down που πραγματώνεται.



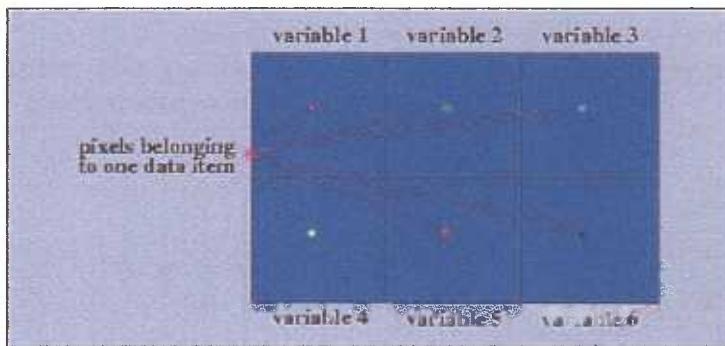
Εικόνα 22

Τεχνική Εξερεύνησης Δομής Καταλόγων - FSN (File System Navigation)

Η τεχνική αυτή είναι σχεδιασμένη από την Silicon Graphics για την εξερεύνηση της δομής των καταλόγων των αποθηκευτικών συσκευών. Χρησιμοποιεί τεχνολογία τριών διαστάσεων και αναπαριστά τους καταλόγους με ένα σύνολο συρταριών όπου κάθε ένα από αυτά περιέχει σύνολα από κουτιά τα οποία αναπαριστούν τα αρχεία. Το μέγεθος του κάθε κουτιού παρουσιάζει το φυσικό μέγεθός του και το χρώμα του παρουσιάζει την ηλικία του. Η FSN χρησιμοποιεί την τεχνική 'artificial perspective'.

2.3.5 Τεχνικές Σχεδίασης σε Επίπεδο Pixels

Η τεχνική αυτή συνίσταται στην παρουσίαση κάθε τιμής ενός γνωρίσματος ενός πίνακα με ένα pixel στην οθόνη χρησιμοποιώντας ένα συγκεκριμένο χρωματισμό, ο οποίος παρουσιάζει τη σχετικότητα (relevance) των δεδομένων. Κάθε γνώρισμα του πίνακα παρουσιάζεται σε διαφορετικά υπό-παράθυρα στην οθόνη



Εικόνα 23

(Εικόνα 23).

Οι τεχνικές αυτές κατηγοριοποιούνται σε 3 είδη [19][21]:

- ✓ **Query dependent** που παρουσιάζουν τα δεδομένα βασιζόμενες σε ερωτήσεις του χρήστη.

Οι αλγόριθμοι, οι οποίοι χρησιμοποιούνται χωρίζονται σε

1. Snake Spiral
2. Snake Axes

- ✓ **Query independent** που απλώς παρουσιάζουν τα δεδομένα. Για να λειτουργήσουν τέτοιου είδους τεχνικές θα πρέπει να υπάρχει μέσα στα δεδομένα μια λογική ταξινόμησης π.χ. time series. Αλγόριθμοι για αυτή την κατηγορία είναι οι εξής:

1. Screen Filling Curve
2. Recursive pattern

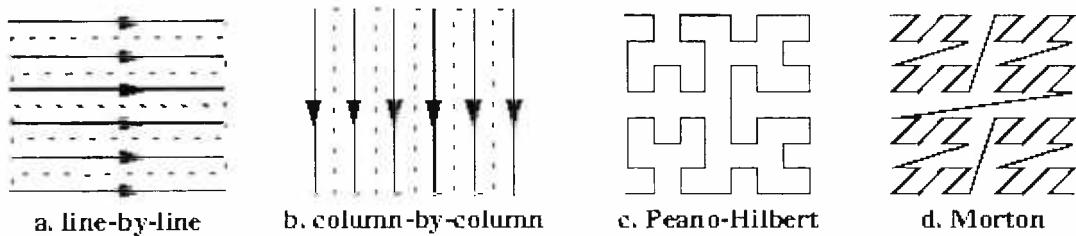
- ✓ **Grouping techniques:** Βασική ιδέα είναι η συνένωση των πολλών παραθύρων που χρησιμοποιούνται για κάθε μεταβλητή σε ένα παράθυρο, ομαδοποιώντας τα δεδομένα για κάθε data item, δηλ για κάθε ξεχωριστή τιμή κάθε στήλης.

Coloring: σε όλες τις παραπάνω τεχνικές χρησιμοποιούνται αλγόριθμοι προσδιορισμού χρωμάτων (HSI) των κουκίδων (π.χ. high = yellow, middle = green, ...).

Παρακάτω αναλύονται οι προαναφερθείσες τεχνικές:

Τεχνικές Ανεξάρτητες Ερωτήσεων (Query Independent)

Ως βασική ιδέα είναι η παρουσίαση όσο το δυνατόν περισσότερων δεδομένων στην οθόνη. Δεν χρησιμοποιούνται τεχνολογίες τρισδιάστατων γραφικών, ενώ αντίθετα το γραφικό αποτέλεσμα έχει μόνο μία διάσταση. Για τον προσδιορισμό του τρόπου τοποθέτησης των pixels στη διάσταση αυτή έχουν σχεδιαστεί διάφοροι αλγόριθμοι, από τους οποίους οι πιο γνωστοί είναι:

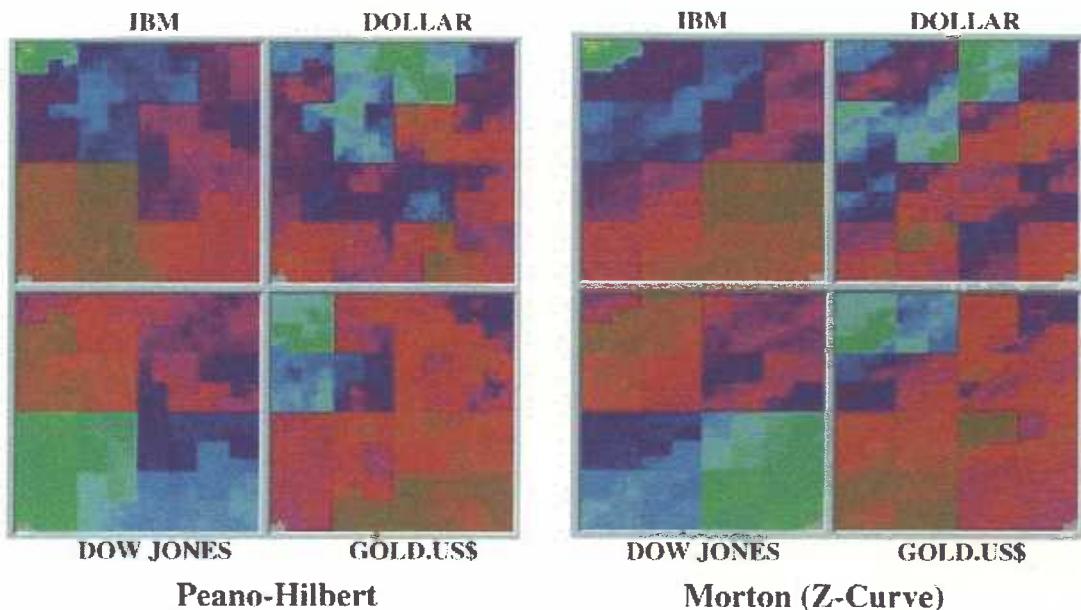


Εικόνα 24

- **Γραμμή - Γραμμή (Line by Line):** Σαρώνεται το αρχείο και τα δεδομένα τοποθετούνται σαν pixels στην οθόνη σειριακά από αριστερά προς τα δεξιά και στη συνέχεια στην επόμενη γραμμή από δεξιά προς τα αριστερά κ.ο.κ. (Εικόνα 24). Για να έχει νόημα αυτού του είδους η παρουσίαση θα πρέπει να έχει προηγηθεί ταξινόμηση των δεδομένων κατά κάποιο γνώρισμα, το οποίο να δίνει ακολουθιακή σημασία στα δεδομένα, π.χ. γνώρισμα-στήλη που περιέχει χρονική πληροφορία, οπότε το αποτέλεσμα θα παρουσίαζε την εξέλιξη των τιμών κάποιας μεταβλητής σε σχέση με το χρόνο.
- **Στήλη - στήλη (Column by Column):** ισχύει ότι και στον αλγόριθμο Γραμμή-Γραμμή με μόνη διαφορά την κατεύθυνση γεμίσματος της οθόνης.
- **Screen Filling Techniques (Peano-Hilbert/Morton):** Βασίζονται στον αλγόριθμο γεμίσματος κενού χώρου του Peano-Hilbert & Morton. Βασική τους χρησιμότητα είναι να παρουσιάζουν δεδομένα, τα οποία εμπεριέχουν έννοια 2 διαστάσεων σε μια διάσταση, περιορίζοντας το χρόνο απεικόνισης και τις απαιτήσεις σε πόρους. Γεμίζουν τετράγωνα διαστάσεων ($2^i \times 2^i$), $i=0..max$. Σε κάθε τέτοιο τετράγωνο υπάρχουν 4 υπό-τετράγωνα μεγέθους ($2^{i-1} \times 2^{i-1}$) και αυτό ισχύει επαναληπτικά. Στον αλγόριθμο Morton (γνωστός και ως Z-curve αλγόριθμος) το γέμισμα έχει μια λογική προσανατολισμού σε αντίθεση με τον αλγόριθμο Peano-Hilbert, δηλ. είναι σαφής η κατεύθυνση με την οποία τοποθετούνται τα Pixels στην οθόνη.

Η τεχνική Peano-Hilbert δεν χρησιμεύει στην ανάγνωση και ερμηνεία των αποτελεσμάτων. Αυτό οφείλεται στο ότι για την παρουσίαση της πληροφορίας ακολουθεί μια διαδρομή τοποθέτησης των pixels στην οθόνη, η οποία δεν είναι εύκολα αναγνώσιμη, ακόμη και όταν είναι γνωστός ο τρόπος σχεδιασμού της. Επίσης εξαιτίας της μη ύπαρξης προσανατολισμού στα αποτελέσματα είναι αδύνατη η σύγκριση αποτελεσμάτων, και κατά συνέπεια,

Space-Filling Curve Arrangements



Εικόνα 25

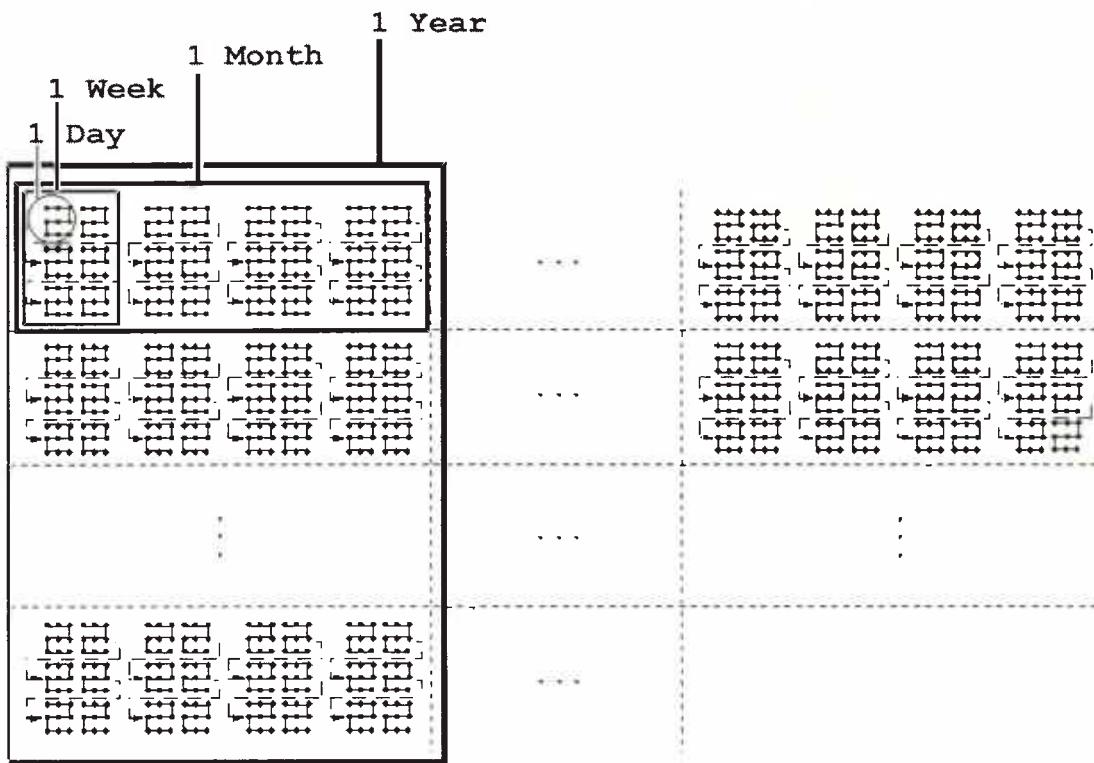
δεν καταφέρνουμε να λάβουμε τη σχετική πληροφόρηση. Το μειονέκτημα αυτό λύνεται με χρήση της τεχνικής Morton, η οποία έχει σαφώς ευκολονόητη ακολουθία γεμίσματος των υπό-τετραγώνων. Στο παράδειγμα της εικόνας 25 παρουσιάζονται δεδομένα που περιγράφουν τις πωλήσεις της IBM, την πορεία του δολαρίου, τον δείκτη Dow Jones και την πορεία του χρυσού. Παρατηρείται ότι το αποτέλεσμα γίνεται δύσκολα κατανοητό (και με τις δύο μεθόδους) όταν εξετάζεται η κάθε μεταβλητή ξεχωριστά. Ωστόσο, η τεχνική Morton δίνει δυνατότητα για συγκριτικά αποτελέσματα μεταξύ των 4 διαφορετικών μεταβλητών, μολαταύτα πρέπει πάλι να γνωρίζεις τον αλγόριθμο για να μην καταλήξεις σε λάθος συμπεράσματα.

- Recursive Pattern Technique:** Ο αλγόριθμος αυτός [20] αποτελεί μια εξελιγμένη μορφή των τεχνικών "Screen Filling". Η βασική ιδέα συνίσταται στη δημιουργία εικόνας με έμφαση στα clusters ενώ ταυτόχρονα δίνει τη δυνατότητα στο χρήστη να επηρεάσει την διάταξη των pixels στην οθόνη ώστε το αποτέλεσμα να είναι πιο κατανοητό. Αυτό γίνεται εμφανές σε δεδομένα με ακολουθιακή εσωτερική δομή π.χ. time series. Η παραπάνω τεχνική βασίζεται σε επαναλαμβανόμενο σχήμα, δηλ. γεμίζει την οθόνη χρησιμοποιώντας επαναλαμβανόμενα σχήματα όπως και

στους screen filling αλγόριθμους, σε σχήμα: αριστερά → δεξιά → επόμενη γραμμή → δεξιά → αριστερά → κ.ο.κ..

Στην ουσία, ο αλγόριθμος αυτός παρουσιάζει τα αποτελέσματα στην οθόνη κατατμίζοντάς την ώστε κάθε τμήμα της, να παρουσιάζει τα δεδομένα με βάση την μεταβλητή ταξινόμησης, π.χ. τον χρόνο. Ταυτόχρονα δημιουργεί ομαδοποιήσεις των δεδομένων, π.χ. όταν ένα υπό-τμήμα της οθόνης παρουσιάζει δεδομένα μίας εβδομάδας τότε τέσσερα υπό-τμήματα μαζί παρουσιάζουν ένα μήνα κ.ο.κ. Κατά αυτόν τον τρόπο, δίνεται iεραρχικό νόημα στο αποτέλεσμα.

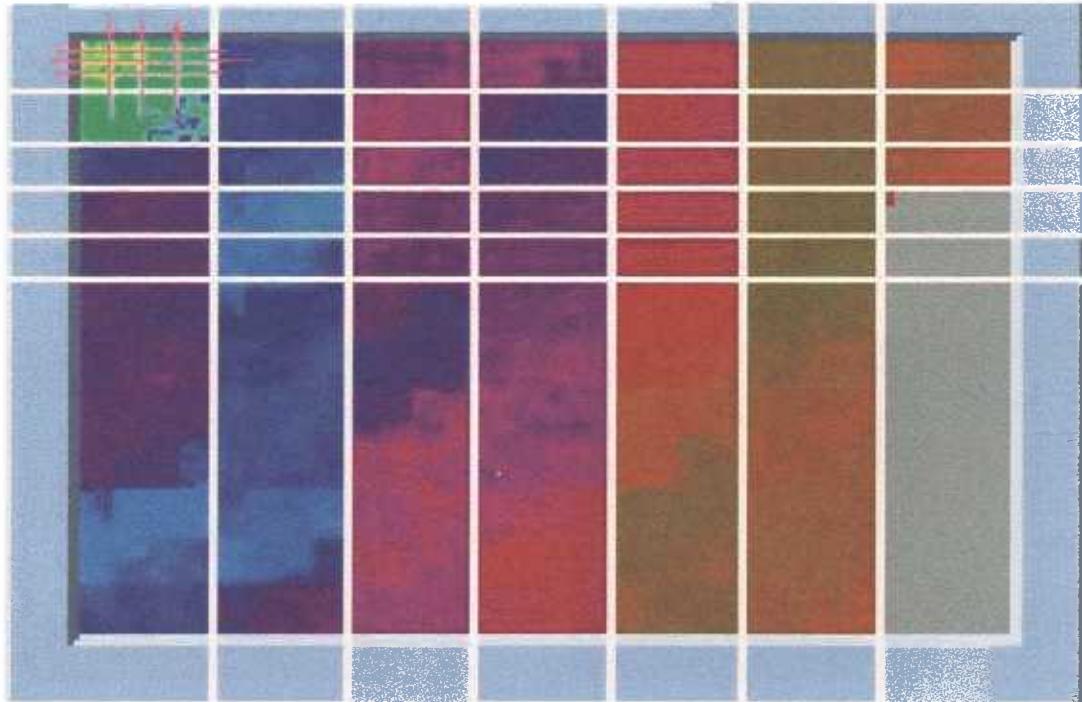
Συγκεκριμένα, ορίζει 2 μετρικές: την w : πλάτος γραμμής και την h : αριθμός γραμμών. Θέτει το διάνυσμα (w_i, h_i) το οποίο ορίζει έναν κύβο μέσα στον οποίο θα τοποθετηθούν $w \times h$ pixels από τα οποία το καθένα αναπαριστά ένα record. Δημιουργεί μια **ιεραρχική λογική** θέτοντας $i=1..max$, όπου το i ορίζει επίπεδα. Το i επίπεδο αποτελείται από $(w_i \times h_i)$ $i-1$ υπό-επίπεδα. Ορίζοντας τα w και h για κάθε επίπεδο, δημιουργείται μια ομαδοποίηση των δεδομένων όπου ο μεγάλος κύβος περιέχει πολλούς μικρότερους κ.λ.π. Τέλος αν οι μετρικές είναι ίσες με τη ρίζα του αριθμού των records τότε δημιουργείται ένα τετράγωνο που απλώς γεμίζει με pixels χωρίς καμία κατηγοριοποίηση.



Εικόνα 26

- Παράδειγμα:** έστω, ένα σύνολο δεδομένων το οποίο παρουσιάζει την μεταβολή του γνωρίσματος x ως προς το χρόνο (Εικόνα 26). Παρατηρώντας την εικόνα βλέπουμε ότι εσωτερικά του γραφήματος χρησιμοποιείται ο αλγόριθμος του

Morton. Η παραπάνω γραφική παρουσίαση χρησιμοποιεί 5 επίπεδα δηλ. $i=1..5$, και οι τιμές των μετρικών είναι οι ακόλουθες: $(w_1, h_1)=(3,3)$, $(w_2, h_2)=(2,3)$, $(w_3, h_3)=(4,1)$, $(w_4, h_4)=(1,12)$, $(w_5, h_5)=(7,1)$. Για το πρώτο επίπεδο έχει τεθεί $(w_1, h_1)=(3,3)$, το οποίο σημαίνει ότι θα χρησιμοποιηθούν 3 επί 3 =9 pixels για την αναπαράσταση 9 εγγραφών οι οποίες αποτελούν τις εγγραφές μίας ημέρας. Στη συνέχεια, το δεύτερο επίπεδο ορίζει τις μετρικές $(w_2, h_2)=(2,3)$, το οποίο σημαίνει ότι τοποθετούμε 2 φορές στον οριζόντιο άξονα τα 3x3 pixels και 3 φορές στον κάθετο, αντλώντας από τη βάση (η οποία είναι ταξινομημένη με βάση το χρόνο) διαδοχικές εγγραφές. Αυτό το βήμα θα δημιουργήσει ένα τμήμα στην οθόνη το οποίο αντιστοιχεί σε βδομάδα. Συνεχίζοντας με την ίδια λογική στα επόμενα



Εικόνα 27

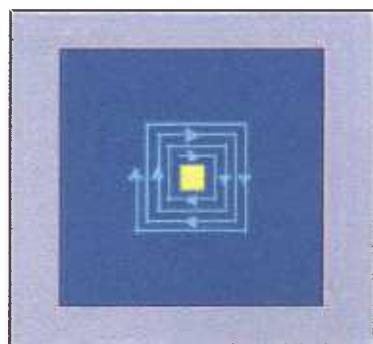
επίπεδα, συνθέτουμε πληροφορία ομαδοποιημένη ανά χρονική περίοδο (ημέρα, εβδομάδα, μήνας, χρόνος κ.λ.π.). Ουσιαστικά δημιουργήθηκε ένα γράφημα με δύο διαστάσεις (χρόνος / παρουσιαζόμενη μεταβλητή με χρήση χρωμάτων) παρουσιαζόμενο σε μία. Το γράφημα που δημιουργείται από χρήση πραγματικών δεδομένων φαίνεται στην εικόνα 27.

Τεχνικές Εξαρτώμενες από Ερωτήσεις (Query Dependent)

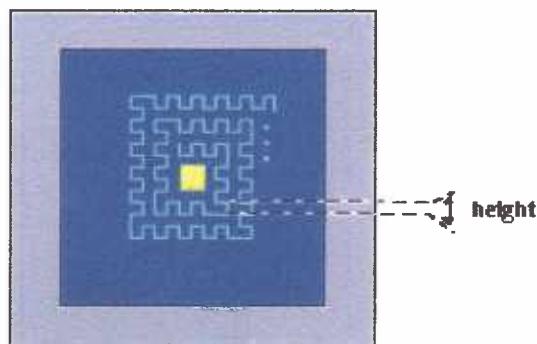
Η τεχνική αυτή σχετίζεται σε μεγάλο βαθμό με την προαναφερθείσα. Το σημείο διαφοροποίησής της είναι ότι δεν παρουσιάζει τα δεδομένα στο σύνολό τους με βάση την μεταξύ τους απόσταση, αλλά με βάση την απόσταση από κάποια ερώτηση που έχει τεθεί από τον χρήστη. Για κάθε στοιχείο κάθε εγγραφής υπολογίζεται η απόστασή από την ερώτηση αυτή, στη συνέχεια υπολογίζεται η συνολική απόσταση των δεδομένων, πάντα με βάση την ερώτηση που τέθηκε και στο

τέλος με βάση τη θέση των δεδομένων στην συνολική απόσταση που υπολογίστηκε, τοποθετούνται τα pixels σε ένα σπειροειδή σχηματισμό γύρω από το κέντρο με τα πιο σχετικά στοιχεία (δηλ. μικρότερη απόσταση από την ερώτηση) πιο κοντά στο κέντρο. Με την τεχνική αυτή ανακαλύπτονται συσχετίσεις, αλληλεξαρτήσεις, συναρτησιακές εξαρτήσεις κ.λ.π.. Συγκεκριμένα, το χρώμα του κάθε pixel προσδιορίζεται από την απόστασή (distance) του από την ερώτηση του χρήστη και η θέση του pixel στην οθόνη προσδιορίζεται με τον ίδιο τρόπο δημιουργώντας μια spiral (ή snake-spiral) τεχνική τοποθέτησης των αντικειμένων γύρω από το κέντρο. Υπάρχουν δύο αλγόριθμοι που υλοποιούν την τεχνική αυτή:

- **Snake-Spiral Technique:** Το 1% των δεδομένων (που είναι πιο κοντά στην ερώτηση του χρήστη) τοποθετούνται στο κέντρο της οθόνης και τα υπόλοιπα δεδομένα γεμίζουν την οθόνη περιμετρικά. Καθώς έχουν υπολογιστεί οι αποστάσεις των δεδομένων από τη σωστή απάντηση, τα δεδομένα, τα οποία απέχουν λιγότερο τοποθετούνται πιο κοντά στο κέντρο (distance metrics). Για κάθε μεταβλητή δημιουργείται ένα υπό-παράθυρο στην οθόνη, και έτσι μπορούμε να έχουμε συγκριτικά αποτελέσματα. Η διαφορά της spiral από τη snake spiral τεχνική είναι ότι στη δεύτερη εμφανίζονται καλύτερα τα clusters που μπορεί να υπάρχουν στα δεδομένα (Εικόνα 28).



a. Spiral Technique



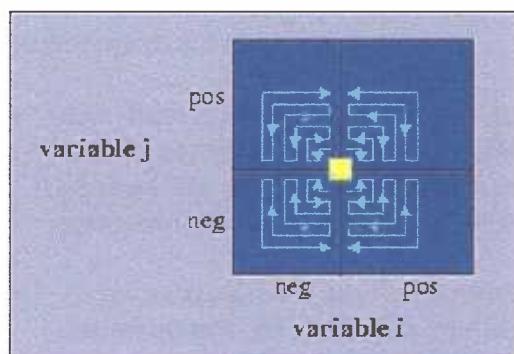
b. Snake-Spiral Technique

Εικόνα 28

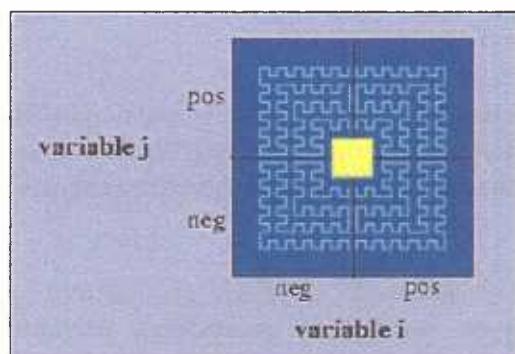
- **Snake-Axes Technique:** Στον αλγόριθμο αυτό χρησιμοποιούνται δύο μεταβλητές. Η βασική ιδέα είναι να δοθεί έμφαση στις κατευθύνσεις των αποστάσεων των τιμών των μεταβλητών από την ερώτηση του χρήστη. Συγκεκριμένα, θέτουμε στον α όνα x την μεταβλητή i και στον α όνα y την μεταβλητή j , οι θετικές τιμές της μεταβλητής i τοποθετούνται στο δεξιό τμήμα της οθόνης και οι αρνητικές στο αριστερό ενώ για τη μεταβλητή j οι θετικές τοποθετούνται στο κάτω τμήμα της οθόνης ενώ οι αρνητικές στο πάνω. Η συγκεκριμένη τεχνική βοηθά στο να δοθεί εικόνα για τον καταμερισμό των τιμών του πίνακα σε σχέση με τις δύο μετρούμενες μεταβλητές (Εικόνα 29).

Τεχνικές Ομαδοποίησης (Grouping)

Η λογική, η οποία χρησιμοποιείται εδώ είναι εντελώς διαφορετική από τις προηγούμενες. Δηλαδή, δεν χρησιμοποιείται ένα υπό-παράθυρο για κάθε μεταβλητή αλλά τα items κάθε μεταβλητής ομαδοποιούνται και γεμίζουν από κοινού την περιοχή γύρω από το κέντρο (Εικόνα 30). Η λογική της τεχνικής αυτής μοιάζει λογικά με την προαναφερθείσα "Dimensional Stacking". Σημαντικό μειονέκτημα είναι η αδυναμία απεικόνισης μεγάλων όγκων δεδομένων.



a. Axes Technique



b. Snake-Axes Technique

Εικόνα 29

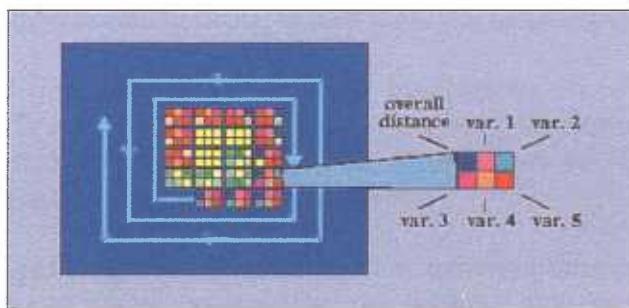


Figure 16: Grouping Technique of Five-Variate Data

Εικόνα 30

Πλεονεκτήματα: Σημαντικό πλεονέκτημα των pixel oriented τεχνικών είναι η απεικόνιση μεγάλων όγκων δεδομένων, μιας και ο χώρος που καταλαμβάνει το κάθε pixel (μια εγγραφή) είναι πολύ μικρός. Μπορούν να αναπαρασταθούν γύρω στις 20.000 εγγραφές. Ένα επιπλέον πλεονεκτήμα είναι ότι δεν χρειάζονται ιδιαίτερα δυνατά μηχανήματα - πόροι για την εφαρμογή των τεχνικών αυτών αφού δεν χρησιμοποιούν τεχνολογίες τρισδιάστατων γραφικών. Τέλος, δεν απαιτούν ιδιαίτερο χρόνο για τη δημιουργία των γραφικού αποτελέσματος σχετικά με τα δεδομένα που απεικονίζουν.

Μειονεκτήματα: Εξαρτώνται σε μεγάλο βαθμό από την ανάλυση της οθόνης. Συγκεκριμένα, όσο μεγαλύτερη είναι η ανάλυση της οθόνης τόσο περισσότερα δεδομένα μπορούν να αναπαρασταθούν. Επίσης, δεν δημιουργούν ιδιαίτερα ευνόητο αποτέλεσμα, γεγονός που τις καθιστά άχρηστες στο μεγαλύτερο μέρος των χρηστών.

2.3.6 Τεχνικές Γραφημάτων (Graph Based)

Μία τελευταία κατηγορία γεωμετρικών τεχνικών είναι τα γραφήματα. Η έννοια γράφημα χρησιμοποιήθηκε πολλές φορές στις μέχρι τώρα αναφερθείσες τεχνικές αλλά δεν πρέπει να συγχέεται με τις τεχνικές που θα περιγραφούν σ' αυτό το κεφάλαιο. Οι τεχνικές, οι οποίες βασίζονται σε γραφήματα έχουν ορισμένα χαρακτηριστικά [1][4] που τις διαχωρίζουν από τις υπόλοιπες. Τέτοια χαρακτηριστικά είναι:

- Δεν παρουσιάζουν πληροφορία σε σύνολα αξόνων. Έχουν τη δυνατότητα να αναπαραστήσουν πολυδιάστατη πληροφορία αλλά χωρίς τη χρήση τρισδιάστατων αξόνων και αυτό τις διαχωρίζει από τις γεωμετρικές τεχνικές με τις οποίες θα μπορούσε κάποιος να τις συσχετίσει.
- Αντικατοπτρίζουν την πραγματική έννοια της λέξης "Γράφημα", δηλ. αποτελούνται από σύνολα γραμμών-ακμών (ευθέων ή καμπύλων) τα οποία ενώνουν σημεία-κόμβους, σχηματίζοντας έτσι πολυγωνικούς σχηματισμούς γραμμών.
- Απευθύνονται κυρίως σε δεδομένα που εμπεριέχουν πληροφορίες συσχέτισης και αλληλεξάρτησης (relational information). Μέσω των γραμμών - ακμών που σχηματίζονται ανάμεσα στους κόμβους δηλώνουν αυτές τις σχέσεις.
- Υπάρχουν γραφήματα 2 και τριών διαστάσεων. Η εισαγωγή τρισδιάστατης τεχνολογίας πραγματώνεται σε ήδη υπάρχουσες δισδιάστατες τεχνικές με σκοπό την καλύτερη εκμετάλλευση χώρου, χωρίς αυτό να αλλοιώνει την υφή του αποτελέσματος (λογική και σημασιολογική).
- Ένα δευτερεύον χαρακτηριστικό είναι η αποφυγή διασταύρωσης των γραμμών, δηλ. επιλέγεται πάντα μια διαδρομή των γραμμών που απεικονίζονται τέτοια ώστε να μην διασταυρώνονται μεταξύ τους αλλά και να μην επικαλύπτουν τα παρουσιαζόμενα σημεία-κόμβους.
- Τα γραφήματα δίνουν μεγάλη έμφαση σε ένα παράγοντα, οποίος δεν επισημάνθηκε ιδιαίτερα από τις άλλες τεχνικές. Ο παράγοντας αυτός ονομάζεται "Αισθητική" και σημαίνει τις λεπτές αλλά σημαντικές αρχές που καθορίζουν τη σχεδίαση ενός γραφήματος. Η αλλαγή αυτών των κανόνων αισθητικής οδηγεί σε διαφορετικές προσεγγίσεις και κατά συνέπεια δημιουργεί νέους τύπους γραφημάτων. Τέτοιοι κανόνες είναι:
 - ⇒ Ο τύπος της ακμής μπορεί να είναι ευθείες, πολυγωνικές γραμμές, καμπύλες κ.λ.π.
 - ⇒ Το μήκος των ακμών μεταβάλλεται ή όχι. Για παράδειγμα όταν το μήκος των ακμών μεταβάλλεται ακαθόριστα μεταξύ των συνδέσεων, κατά συνέπεια και οι κόμβοι δεν ισαπέχουν μεταξύ τους. Αν ισχύει ή όχι αυτός ο κανόνας έχει σαν συνέπεια παραγωγή διαφορετικού τύπου γραφήματος.



- ❖ Το μέγεθος (ύψος, πλάτος) και ο χρωματισμός των κόμβων.
- ❖ Ύπαρξη ομοιόμορφης κατανομής των ακμών στο διαθέσιμο χώρο.
- ❖ Η δημιουργία συμμετρικού ή ισοζυγισμένου γραφήματος.

Αν και από την πλευρά απεικόνισης εμφανίζεται ως μία εύκολη τεχνική, στην ουσία κρύβει πολλές δυσκολίες και παγίδες, όπως ενδεικτικά οι ακόλουθες:

- Χρειάζονται πολύπλοκοι αλγόριθμοι για να υπολογίζουν τον τρόπο σχεδίασης των ακμών ώστε να μην περιπλέκονται, αποφεύγοντας έτσι ένα ακατανόητο αποτέλεσμα.
- Χρειάζονται ειδικοί αλγόριθμοι που θα υπολογίζουν την κατανομή και τα σημεία τοποθέτησης των κόμβων στην οθόνη.
- Χρειάζεται να ελέγχεται ο βαθμός καμπυλότητας των ακμών
- Όλα τα παραπάνω πρέπει να υπολογίζονται σε αποδεκτό χρόνο

Στη συνέχεια παρουσιάζονται 4 τεχνικές, οι οποίες ανήκουν στην κατηγορία των δισδιάστατων γραφημάτων. Από την παρουσίαση αυτή γίνεται σαφές ότι ο τρόπος, με τον οποίο η κάθε τεχνική θεσπίζει τους κανόνες αισθητικής, "ορίζει" και τον τύπο της. Δηλ. στην πραγματικότητα, δεν αλλάζει η λογική απεικόνισης, αλλά αλλάζουν οι κανόνες που προαναφέρθηκαν.

Ορθογώνιο Γράφημα

Οι ακμές είναι είτε ευθείες είτε πολυγωνικές με ορθογώνιες γωνίες. Στο γράφημα της εικόνας 31 χρησιμοποιούνται 3 διαστάσεις οι οποίες απεικονίζονται από το ύψος, πλάτος και χρώμα του κόμβου, ενώ ταυτόχρονα παρουσιάζεται και η συσχέτιση ή η αλληλεξάρτηση των κόμβων.

Συμμετρικό Γράφημα

Βασικής σημασίας στην τεχνική αυτή είναι η συμμετρική εικόνα του γραφήματος. Οι κόμβοι είναι έτσι κατανεμημένοι ώστε να παράγεται συμμετρικό αποτέλεσμα (Εικόνα 32). Η διαδικασία αυτή απαιτεί πολλούς μαθηματικούς υπολογισμούς δεδομένου ότι η τοποθέτηση των κόμβων γίνεται με βάση τις υπάρχουσες συσχετίσεις, ώστε η διασύνδεσή τους μέσω των ακμών να μην περιπλέκει το γράφημα.

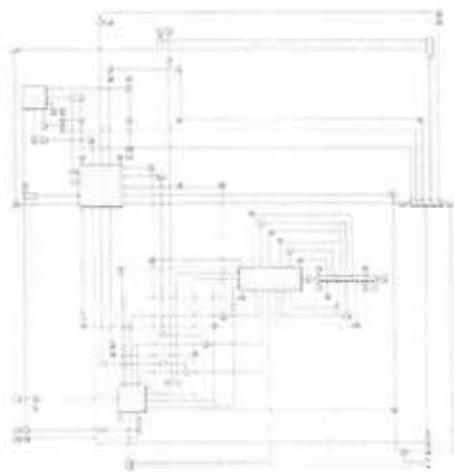
Γράφημα Κατηγοριοποίησης (Cluster Based)

Στόχος του αλγορίθμου της τεχνικής είναι να βρει τους κόμβους που παρουσιάζουν τις περισσότερες συνδυαστικές αλληλοσυσχετίσεις, και να τους τοποθετήσει σε κάποιο κοινό σημείο ώστε να σχηματίσουν cluster (Εικόνα 33).

Κατευθυνόμενο - ακυκλικό γράφημα (Ayclic Graph)

Ο κεντρικός άξονας της τεχνικής αυτής είναι η δημιουργία ενός γραφήματος όπου οι ακμές έχουν την έννοια της κατευθυνσης. Επίσης, δεν επιτρέπεται η





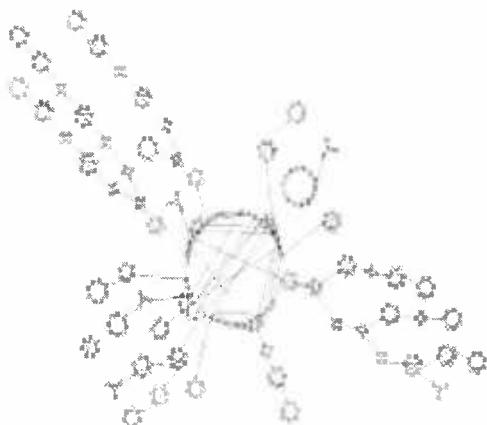
Orthogonal Graph

Εικόνα 31



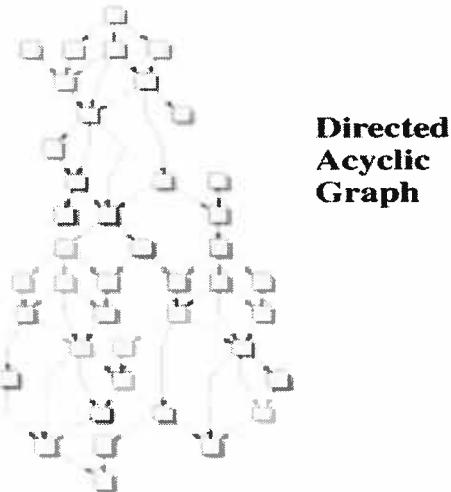
Symmetry-Optimized Graph

Εικόνα 32



Cluster-Optimized Graph

Εικόνα 33



Εικόνα 34

δημιουργία κυκλικών δομών μεταξύ των κόμβων (Εικόνα 34). Αυτό σημαίνει ότι για κάθε κόμβο δεν υπάρχει κανένα μονοπάτι που να οδηγεί στον εαυτό του. Η συγκεκριμένη δομή είναι ιδιαίτερα πρόσφορη για παρουσίαση δεδομένων, τα οποία εμπεριέχουν ιεραρχική πληροφορία.

Για όλες τις παραπάνω τεχνικές γραφημάτων έχουν πραγματοποιηθεί προσπάθειες απεικόνισής τους στον τρισδιάστατο χώρο, με στόχο την καλύτερη εκμετάλλευση του διαθέσιμου χώρου μίας οθόνης υπολογιστή.

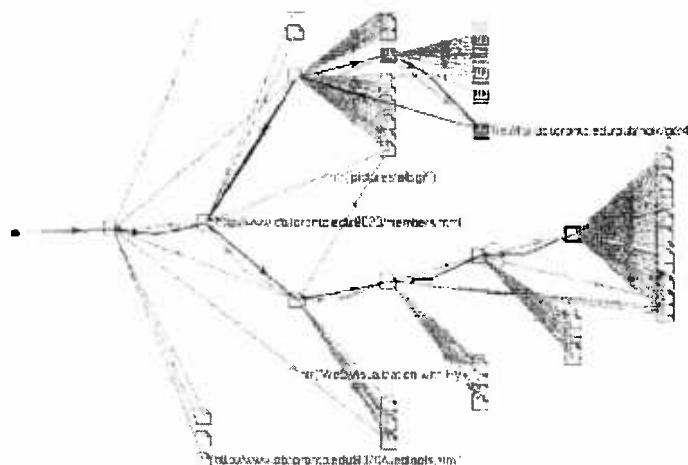
Στη συνέχεια παρουσιάζονται μερικές από τις πιο ανεπτυγμένες τεχνικές γραφημάτων, οι οποίες χρησιμοποιούν τεχνολογία τρισδιάστατων γραφικών και πολυδιάστατης παρουσίασης.

Γραφήματα Υπέρ-Δεντρικών Λομών (Hygraphs)

Τα Hygraphs (Εικόνα 35) εμπλουτίζουν τη λειτουργικότητα των κόμβων προσδίδοντάς τους μία πολυδιάστατη εικόνα. Συγκεκριμένα τα Hygraphs αποτελούνται από τις γνωστές ακμές οι οποίες δηλώνουν τις συσχετίσεις των δεδομένων και από μία νέα μορφή κόμβων οι οποίοι περιέχουν 0 ή περισσότερους υπό-κόμβους. Επιλέγοντας ένα σύνθετο κόμβο, ανοίγεται ένα νέο σύνολο δεδομένων με ακμές και κόμβους, οι οποίοι μπορεί να είναι επίσης σύνθετοι. Τα αποτελέσματα αυτής της κατηγοριοποίησης είναι:

- Κατηγοριοποίηση των δεδομένων, μιας και όλοι οι κόμβοι που ανήκουν σε ένα σύνθετο κόμβο παρουσιάζουν μια αλληλεξάρτηση με αυτόν.
- Καλύτερη εκμετάλλευση του χώρου της οθόνης, αφού δεν παρουσιάζονται ταυτόχρονα όλα τα δεδομένα της βάσης δεδομένων.
- Δυνατότητα δυναμικής αλληλεπίδρασης με το χρήστη, ο οποίος μπορεί να ανοίξει ένα σύνθετο κόμβο επιλέγοντάς τον.
- Αποτελεσματικότερη παρουσίαση των δεδομένων όσον αφορά τον τρόπο με τον οποίο ο χρήστης παρατηρεί και κατανοεί τα δεδομένα. Αφενός, δεν χάνεται ο έλεγχος του οπτικού αποτελέσματος αφού περιορίζεται το μέγεθος των παρουσιαζόμενων δεδομένων και αφετέρου, παρουσιάζονται καλύτερα οι συσχετίσεις και οι κατηγοριοποήσεις αυτών.
- Δυνατότητα χρήσης εικονιδίων και λεκτικών για αναπαράσταση των κόμβων

Η τεχνική αυτή χρησιμοποιείται κυρίως για αναπαράσταση ιεραρχικής πληροφόρησης, όπως για παράδειγμα η δομή των καταλόγων ενός δίσκου, ή ακόμη καλύτερα η δομή ενός καταλόγου που εμπεριέχει WEB σελίδες. Στην περίπτωση αυτή οι σύνδεσμοι (Hyperlinks) από μία σελίδα σε μία άλλη θα μπορούσαν να παρουσιάζονται μέσω των ακμών ενώ η ιεραρχική δομή των καταλόγων μπορεί να παρουσιάζεται μέσω των σύνθετων κόμβων. Τέλος οι ακμές (οι WEB σύνδεσμοι για το συγκεκριμένο παράδειγμα) μπορούν να είναι κατευθυνόμενες, ενώ οι ακμές των σύνθετων κόμβων με τα παιδιά τους να μην είναι.



Εικόνα 35

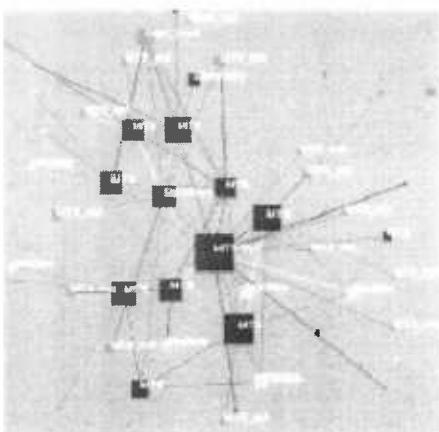
Γραφήματα SeeNet

Η τεχνική αυτή οπτικοποιεί ιεραρχικά δίκτυα με βάρη. Τα σημαντικά χαρακτηριστικά της μεθόδου αυτής είναι η δυνατότητα αναπαράστασης ιεραρχικής πληροφορίας και η προσαρμογή βαρών για αναγνώριση του βαθμού συσχέτισης των δεδομένων. Το δεύτερο χαρακτηριστικό είναι ιδιαίτερα σημαντικό για την χρήση της τεχνικής προκειμένου για την αναπαράσταση των αποτελεσμάτων της διαδικασίας εύρεσης συσχετίσεων (associations) των διαδικασιών εξόρυξης γνώσης. Συγκεκριμένα, τα χαρακτηριστικά της μεθόδου είναι:

- Σημασιολογική θέση των κόμβων. Η τοποθέτηση των κόμβων στο γράφημα δεν πραγματοποιείται τυχαία αλλά σε σχέση με την μεταξύ τους συσχέτιση, δηλ. κόμβοι με μεγάλα βάρη είναι πιο κοντά ο ένας στον άλλον.
- Οι ιδιότητες των δεδομένων απεικονίζονται στο χρώμα και το μέγεθος (ή πάχος) των ακμών και των κόμβων.
- Άλληλεπίδραση με το χρήστη (Drill Down, hilight, rotate κ.λ.π.).

Στην εικόνα 36 απεικονίζεται η κατάσταση του συστήματος αποστολής ηλεκτρονικών μηνυμάτων μετά από παρακολούθηση μερικών ημερών. Οι ιδιότητες των δεδομένων που απεικονίζονται στο γράφημα αυτό είναι οι ακόλουθες:

1. Κόμβος: γραμματοκιβώτιο (mailbox) κάθε ατόμου της εταιρείας.
2. Ακμή: αποστολή μηνύματος από κόμβο σε κόμβο.
3. Μέγεθος κόμβων: αριθμός των e-mails που έχει το κάθε άτομο της εταιρείας.
4. Χρώμα κόμβων: θέση των ατόμων στην εταιρεία.
5. Μέγεθος ακμών: αριθμός μηνυμάτων που έχουν αποσταλεί μεταξύ των κόμβων που συνδέονται. Μεγάλο πλάτος σημαίνει πολλά μηνύματα.
6. Χρώμα ακμών: Μπλε σημαίνει λίγα μηνύματα και κόκκινο πολλά.

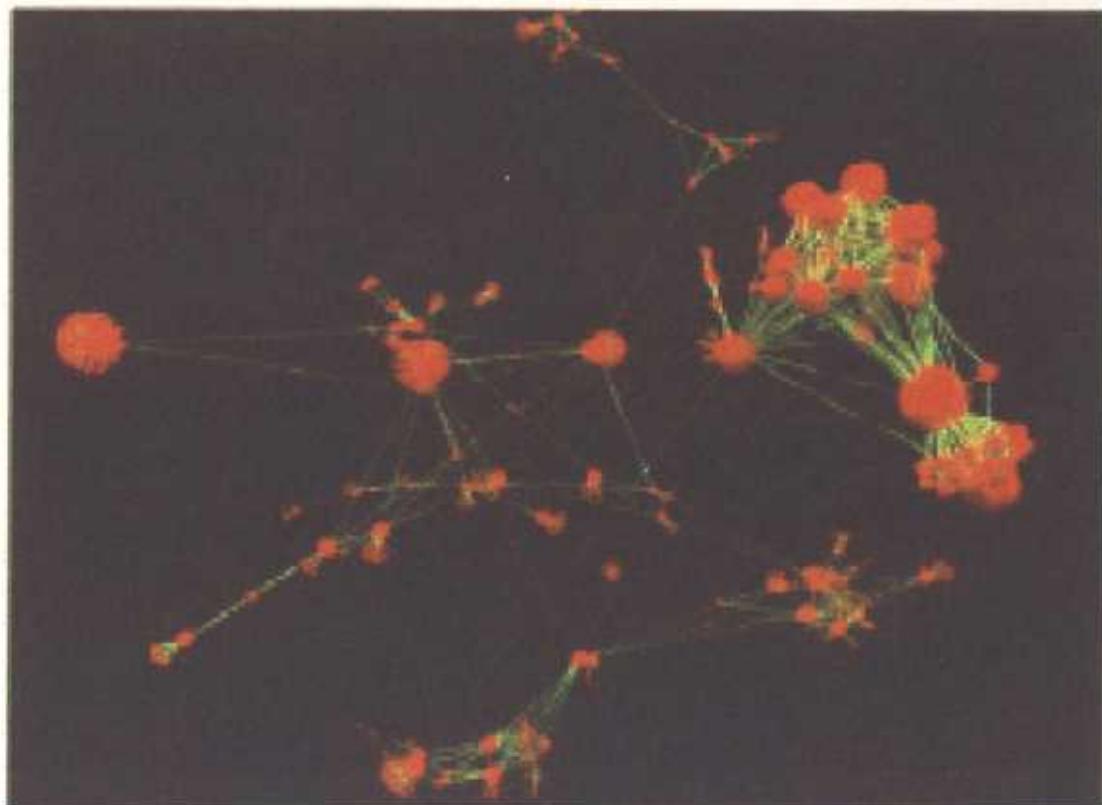


Εικόνα 36

Γράφημα Narcissus

Η τεχνική αυτή χρησιμοποιείται για να απεικονίσει δεδομένα που περιέχουν υψηλό αριθμό συσχετίσεων [31]. Βασικός τομέας εφαρμογής της είναι το Internet (Εικόνα 37), όπου κάθε κόμβος αναπαριστά μία σελίδα και οι ακμές παρουσιάζουν τις συνδέσεις μεταξύ αυτών των σελίδων. Χρησιμοποιείται τεχνολογία τρισδιάστατων γραφικών, και οι ιδιότητες των σελίδων παρουσιάζονται με βάση το χρώμα και το μέγεθος των κόμβων. Με επιλογή του κάθε κόμβου ενεργοποιείται ο εξερευνητής του Internet και αναζητά την επιλεγμένη σελίδα.

Η τεχνική Narcissus θα μπορούσε κάλλιστα να χρησιμοποιηθεί για επεξεργασία των ιστορικών σελίδων που έχει επεξεργαστεί κάθε χρήστης. Επίσης, θα μπορούσε να χρησιμοποιηθεί και από μηχανές αναζήτησης σελίδων ή κειμένων (search agents, text retrieval). Το μειονέκτημα της, όμως, είναι η αδυναμία παρουσίασης αναλυτικής πληροφόρησης για την κάθε σελίδα.



Εικόνα 37



2.3.7 Τεχνικές Παραμόρφωσης Εικόνας (Distortion)

Οι τεχνικές αυτές έχουν ως κύριο στόχο την ελεγχόμενη παραμόρφωση του οπτικού αποτελέσματος προκειμένου να είναι δυνατή η παρουσίαση περισσότερων δεδομένων στην οθόνη. Στις περισσότερες περιπτώσεις, δίνεται ιδιαίτερη έμφαση στα πιο σχετικά, με κάποια ερώτηση του χρήστη, δεδομένα, μειώνοντας τον χώρο απεικόνισης για τα υπόλοιπα. Στην συνέχεια γίνεται μια αναλυτικότερη αναφορά σε ένα σύνολο αντιπροσωπευτικών τεχνικών αυτού του είδους.

Toίχος με Προοπτική (Perspective Wall)

Η τεχνική αυτή απευθύνεται σε δεδομένα, τα οποία έχουν γραμμική δομή [3][29]. Ένα σύνολο δεδομένων που χαρακτηρίζεται από χρονική διαδοχή πράξεων θα μπορούσε να θεωρηθεί ότι περιέχει γραμμική πληροφορία.

Η απεικόνιση γραμμικής πληροφορίας έχει τα ακόλουθα μειονεκτήματα:

1. Όταν το μέγεθος των δεδομένων είναι μεγάλο (π.χ. μεγάλο χρονικό διάστημα συλλογής πληροφοριών) τότε είναι αδύνατο να χωρέσουν όλα τα δεδομένα στην οθόνη σειριακά. Επειδή δεν υπεισέρχεται ιεραρχική πληροφορία, δεν μπορούν αφενός να χρησιμοποιηθούν γνωστές ιεραρχικές τεχνικές, ενώ αφετέρου δεν είναι δυνατή η παρουσίασή τους σειριακά σε μία οθόνη.
2. Δεν είναι εύκολο να προσδιορίσεις μεγάλο aspect ratio, και ταυτόχρονα να ληφθεί ένα κατανοητό αποτέλεσμα.

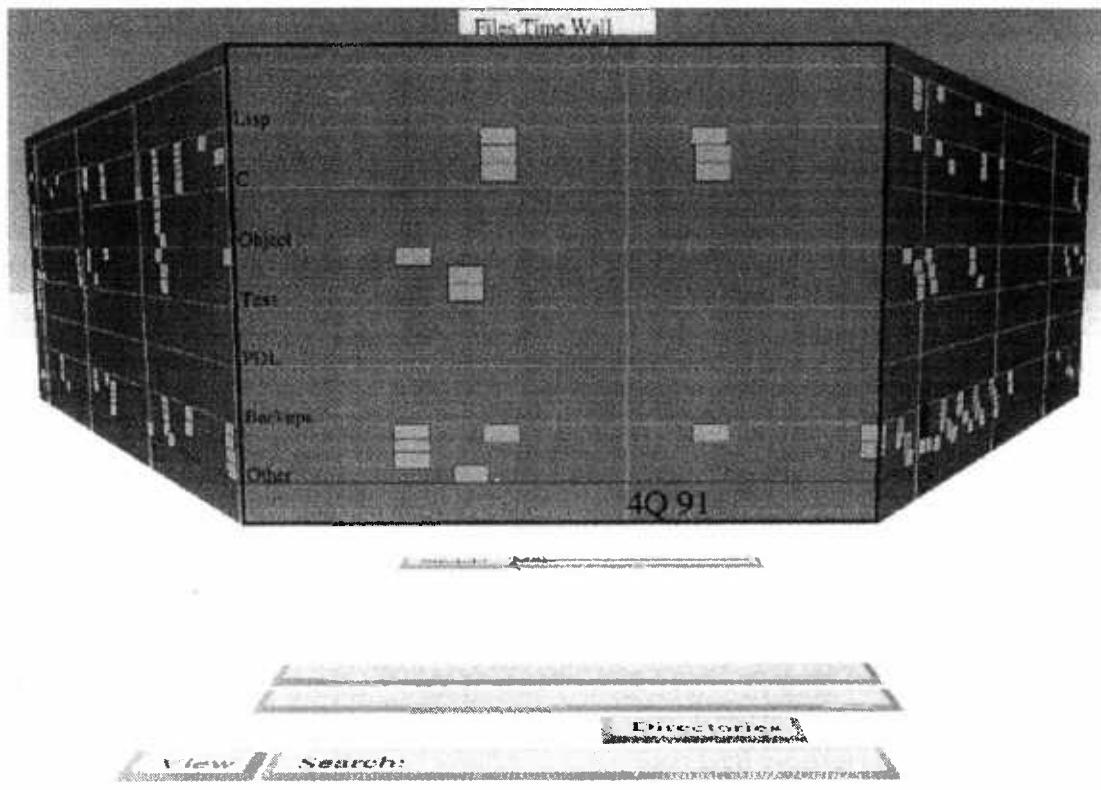
Έτσι, υπάρχει πρόβλημα παρουσίασης του συνόλου της πληροφορίας σε αναλυτικό επίπεδο. Συνεπώς, δεν συμβαδίζει η ολοκληρωμένη εικόνα με την αναλυτική πληροφόρηση.

Μία λύση στο πρόβλημα αυτό είναι η δημιουργία μιας τεχνικής, η οποία να υλοποιεί τη γνωστή λογική "Master Detail". Σ' αυτή την περίπτωση, το σύνολο της πληροφορίας παρουσιάζεται στην οθόνη και όταν επιλεγεί κάποιο συγκεκριμένο τμήμα του τότε υπάρχει ένα μέρος της οθόνης όπου παρουσιάζονται αναλυτικές πληροφορίες για την εκάστοτε επιλογή.

Η τεχνική Perspective Wall, χρησιμοποιώντας την ιδέα της παραμόρφωσης της εικόνας έρχεται να παρουσιάσει το σύνολο της πληροφορίας. Αποτελεί μια εξέλιξη της τεχνικής Bifocal Lens, προσθέτοντας συγχρόνως μία ακόμη διάσταση. Η Bifocal Lens σχεδιάστηκε για μία εταιρεία που συγκέντρωνε πληροφορίες για άρθρα, εφημερίδες, περιοδικά. Τα βασικά αποτελέσματα της ερώτησης του χρήστη παρουσιάζονταν σε μια οριζόντια παράταξη στην οθόνη, ανά σειρά; Δηλ. Οι εφημερίδες καταλάμβαναν μία σειρά, τα περιοδικά την επόμενη κ.λ.π. Στη συνέχεια παρουσιάζονταν και δεδομένα άλλων χρονικών περιόδων που σχετίζονταν με τα αποτελέσματα, στις δύο πλευρές (αριστερά και δεξιά) της οθόνης, εισάγοντας κάποια παραμόρφωση, η οποία γινόταν μεγαλύτερη όσο απομακρύνομασταν χρονικά.



Στην Perspective Wall, υλοποιείται η παραπάνω λογική εισάγοντας τρισδιάστατη τεχνολογία και πολυμέσα (Εικόνα 38). Τα βασικά αποτελέσματα της ερώτησης παρουσιάζονται στο κέντρο της οθόνης σε ένα μεγάλο παραλληλόγραφο που που προσομοιώνει ένα τοίχο, και τα σχετικά αποτελέσματα (στην ουσία οι πιο κοντινές εγγραφές, αφού μιλάμε για γραμμική δομή πληροφορίας) παρουσιάζονται σε δύο τοίχους αριστερά και δεξιά οι οποίοι δημιουργούν την τρισδιάστατη εντύπωση, μιας και μικραίνουν όσο τα δεδομένα απομακρύνονται από τα παρουσιαζόμενα. Στα σχετικά δεδομένα εισάγεται και η τεχνική της παραμόρφωσης.



Εικόνα 38

Το παράδειγμα της εικόνας 38 παρουσιάζει την δομή των αρχείων ενός αποθηκευτικού συστήματος ως προς την ημερομηνία τροποποίησής τους, χρησιμοποιώντας τις γραμμές για να δηλώσει διαφορετικούς τύπους αρχείων. Ακολουθεί μία αναφορά των χαρακτηριστικών της τεχνικής αυτής:

- Όσο απομακρύνονται τα παρουσιαζόμενα δεδομένα από την ενεργή επιλογή, τόσο μειώνεται το μέγεθος του τοίχου αλλά και η αναπαράστασή τους. Αυτό βοηθάει τον χρήστη να καταλάβει μέσω του οπτικού αποτελέσματος ότι μειώνεται η συσχέτιση που υπάρχει ανάμεσα τους.
- Όταν επιλεγεί κάποιος πλαιϊνός τοίχος, τότε αυτός έρχεται στο προσκήνιο με χρήση animation.
- Ο χρήστης έχει τη δυνατότητα να επιλέξει την διαφορά μεγέθους (ratio) μεταξύ του κεντρικού αποτελέσματος και του γενικού πλαισίου. Μπορεί δηλαδή, να

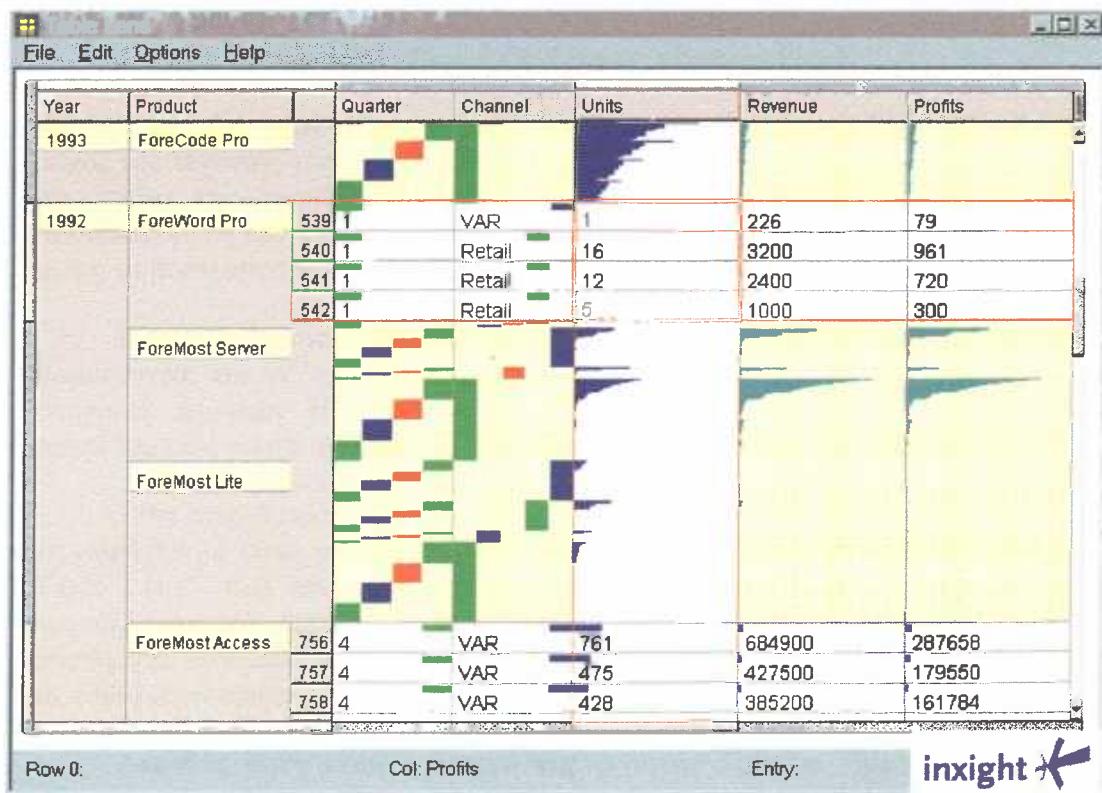
επιλέξει το μέγεθος του κεντρικού παραλληλογράμμου, στο οποίο περιέχονται τα αναλυτικά δεδομένα.

Τέλος η τεχνική αυτή χρησιμοποιείται αποτελεσματικά σε διαδικασίες αναζήτησης δεδομένων ή εγγράφων (text retrieval), όπου στο κεντρικό τμήμα παρουσιάζονται τα αποτελέσματα της ερώτησης και στα πλαϊνά τμήματα παρουσιάζονται όλα τα συσχετιζόμενα στοιχεία - έγγραφα.

Τεχνική Φακού Μεγέθυνσης σε Πίνακες (Table Lens)

Η Table Lens [30] είναι μία τεχνική για παρουσίαση των δεδομένων ενός πίνακα σε μορφή γραφήματος αλλά και αναλυτικών εγγραφών μαζί. Στην ουσία, οι τιμές των στηλών του πίνακα αντικαθιστούνται από την οπτική τους αναπαράσταση σε μία μορφή γραφικής απεικόνισης. Ανάλογα με την πληροφορία που περιέχουν επιλέγεται και ο τρόπος αναπαράστασής τους. Συνήθως, οι μεταβλητές, οι οποίες περιέχουν την πληροφορία που κατηγοριοποιεί τα δεδομένα, ή την πληροφορία που δημιουργεί ιεραρχική δομή, παρουσιάζονται με χρήση των λεκτικών τους, ενώ οι μεταβλητές - στήλες που περιέχουν μετρήσιμη - αριθμητική πληροφορία παρουσιάζονται υπό την μορφή γραφήματος.

Με αυτό τον τρόπο μπορούν να παρουσιαστούν ακόμη περισσότερες εγγραφές μιας και ο τρόπος απεικόνισης των τιμών μπορεί να συμπυκνώσει τα δεδομένα. Έτσι, ο χρήστης μπορεί να πάρει μια εικόνα για το σύνολο των δεδομένων, αφού με μία ματιά βλέπει ένα σύνολο από γραφικές απεικονίσεις των δεδομένων οι οποίες είναι πολύ πιο εύκολα κατανοητές από τα νούμερα ενός πίνακα.



Εικόνα 39

Το στοιχείο διαφοροποίησης της τεχνικής αυτής από τις υπόλοιπες προσεγγίσεις είναι η δυνατότητα αναλυτικότερης διερεύνησης των εγγραφών που παρουσιάζουν ενδιαφέροντα χωρίς απώλεια του συνολικού πλαισίου. Συγκεκριμένα, ο χρήστης έχει την συνολική εικόνα της κατάστασης των δεδομένων του αλλά ταυτόχρονα και χωρίς να χάσει την απεικόνιση αυτή μπορεί να πάρει αναλυτική πληροφόρηση για ένα σύνολο από εγγραφές οι οποίες παρουσιάζουν ιδιόμορφη συμπεριφορά. Για τις εγγραφές που θα επιλεγούν θα αντικατασταθεί το γράφημα από τις εγγραφές που το δημιουργούν σε μορφή λογιστικού φύλου (Εικόνα 39).

Τα πλεονεκτήματα είναι:

- Μεγάλος βαθμός αλληλεπίδρασης με το χρήστη
- Εύκολη μεταλλαγή από συνολικό επίπεδο πληροφόρησης σε πιο λεπτομερειακή ανάλυση
- Παρουσίαση των αποτελεσμάτων με τρόπο ώστε να μην χάνει ο χρήστης την επαφή του με το συνολικό αποτέλεσμα.
- Ολοκληρωμένη μορφή παρουσίασης των δεδομένων (γράφημα, εγγραφές- row data).
- Έμφαση σε τμήμα δεδομένων. Επιλέγοντας ένα συγκεκριμένο σύνολο δεδομένων ο χρήστης επικεντρώνει την προσοχή (κέντρο ενδιαφέροντος) σε αυτά.

Τεχνική με Προοπτική Φακού Μεγέθυνσης (Fisheye View)

Η τεχνική αυτή είναι αρκετά γνωστή και παλιά τεχνική [26](1986). Βασικός στόχος της τεχνικής είναι να επικεντρώνει την παρουσίασή της στα δεδομένα εκείνα που έχουν την μεγαλύτερη σημασία. Ουσιαστικά, θέτονται βάρη για κάθε παρουσιαζόμενη μονάδα δεδομένων, και στη συνέχεια η οπτική απεικόνιση λαμβάνει υπόψη τα βάρη αυτά για να σχεδιάσει τα αντικείμενα.

Η τεχνική αυτή είναι μία επέκταση των τεχνικών γραφημάτων. Εκεί εφαρμόστηκε και γι' αυτά σχεδιάστηκε. Ο λόγος που την εντάξαμε σ' αυτή την κατηγορία τεχνικών είναι επειδή κάνει χρήση της λογικής παραμόρφωσης του αποτελέσματος για τα αντικείμενα που παρουσιάζουν ιδιαίτερο ενδιαφέρον.

Στην πραγματικότητα η τεχνική αυτή χρησιμοποιείται από πολλές άλλες και ένα παράδειγμα είναι οι δύο τεχνικές που προαναφέρθηκαν "Perspective Wall" και "Table Lens". Και στις δύο υπήρχε παραμόρφωση της εικόνας με σκοπό 1.την εκμετάλλευση του διαθέσιμου χώρου και 2.την επικέντρωση της προσοχής του χρήστη στα δεδομένα που έχουν μεγαλύτερο ενδιαφέρον ή στα δεδομένα που είναι πιο κοντά στην ερώτησή του.

Ακριβώς αυτή είναι η λογική της τεχνικής Fisheye. Παρουσίαση της πιο ενδιαφέρουσας πληροφορίας στο κέντρο της οθόνης με εμφανή διαφοροποίηση από



τα άλλα δεδομένα. Όσον, αφορά τα γραφήματα, η επίδειξη του κέντρου ενδιαφέροντος γίνεται μέσα από τον προσδιορισμό του μεγέθους των κόμβων. Οι πιο σημαντικοί κόμβοι παρουσιάζονται με μεγάλο μέγεθος και με μακρύτερες ακμές, ενώ οι μικρότερης σημασίας κόμβοι απεικονίζονται μικρότεροι σε όγκο και πολύ κοντά ο ένας στον άλλο. Έτσι, ο παραπάνω χώρος που χρειάζονται οι μεγάλοι κόμβοι για να απεικονιστούν, κερδίζεται από την μικρότερη απεικόνιση των υπολοίπων κόμβων. Επίσης για τους μεγάλης σημασίας κόμβους παρουσιάζεται κάποια πληροφόρηση μέσω λεκτικών στο εσωτερικό τους. Όσο πιο μεγάλος είναι ο κόμβος τόσο πιο εμφανής είναι η πληροφορία που παρέχει το λεκτικό.

Το σημαντικότερο πλεονέκτημα της τεχνικής αυτής είναι το ακόλουθο: Παρουσίαση του συνολικού πλαισίου των δεδομένων, με ταυτόχρονη λεπτομερέστερη παρουσίαση των σημαντικών κόμβων. Βέβαια η θέση της τεχνικής αυτής σε σχέση με τις δύο παραπάνω τεχνικές είναι μειονεκτική, και αυτό οφείλεται κυρίως στην ηλικία της μεθόδου, μιας και οι παραπάνω μέθοδοι βασίστηκαν στην λογική που η fisheye δημιούργησε. Από την άλλη πλευρά πλεονεκτεί των δύο άλλων τεχνικών στο ότι είναι η πιο κατάλληλη για παρουσίαση ιεραρχικής πληροφόρησης, επειδή το γράφημα μπορεί εύκολα να μετατραπεί σε ιεραρχικό δέντρο.

Στο παρακάτω παράδειγμα είναι εμφανής η υπεροχή του οπτικού αποτελέσματος της fisheye view (Εικόνα 41) σε ένα απλό γράφημα (Εικόνα 40). Το κέντρο του ενδιαφέροντος προσδιορίζεται με βάση συναρτήσεις υπολογισμού ενδιαφέροντος οι οποίες τροφοδοτούνται με στοιχεία τα οποία είναι αναγκαία και ικανά για τον υπολογισμό του ενδιαφέροντος. Υπάρχουν πολλοί αλγόριθμοι υπολογισμού του ενδιαφέροντος οι οποίοι δίνουν βάρος σε διαφορετικά σημεία. Στο παράδειγμα, το κέντρο του ενδιαφέροντος είναι ο κόμβος που αντιπροσωπεύει την πόλη St. Louis. Το σύνολο του γραφήματος παρουσιάζει όλες τις πόλεις των ΗΠΑ και οι ακμές δηλώνουν τις οδούς που συνδέουν τις πόλεις αυτές.

Τα προβλήματα που παρουσιάζονται σ' αυτές τις μεθόδους επικεντρώνονται κυρίως σε θέματα:

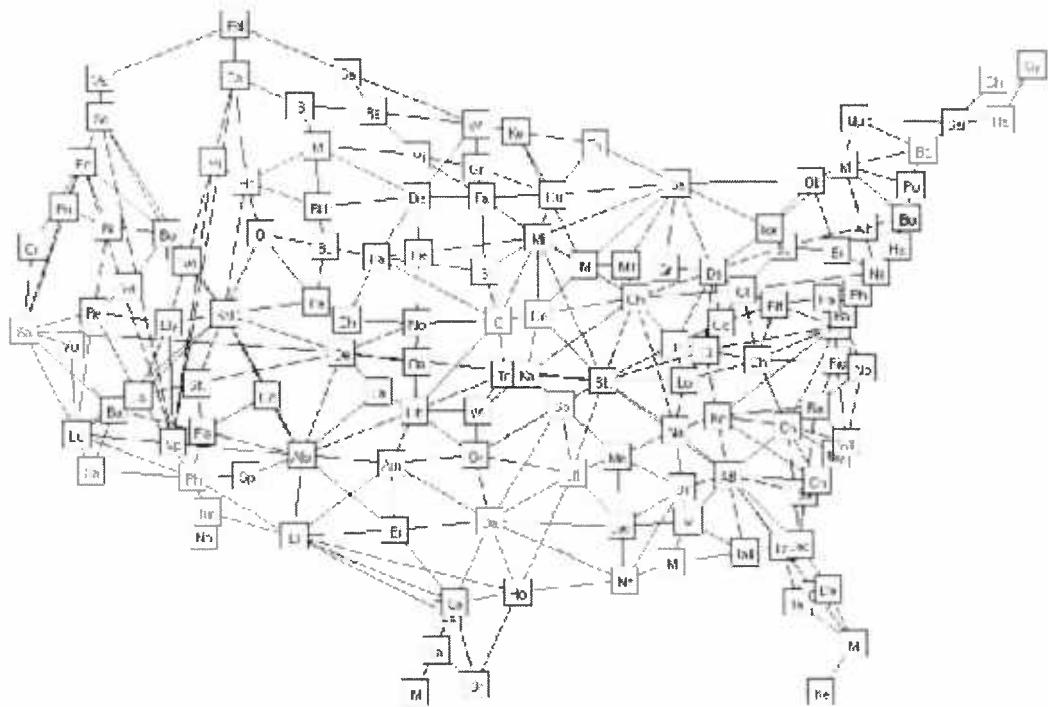
- χρόνου που απαιτούν οι αλγόριθμοι για τον υπολογισμό των κόμβων ενδιαφέροντος και
- επιλογής των σωστών αλγορίθμων ανάλογα με την υφή των δεδομένων.

Τεχνική Δέντρων «Υπερβολικής» Γεωμετρίας (Hyperbolic Trees)

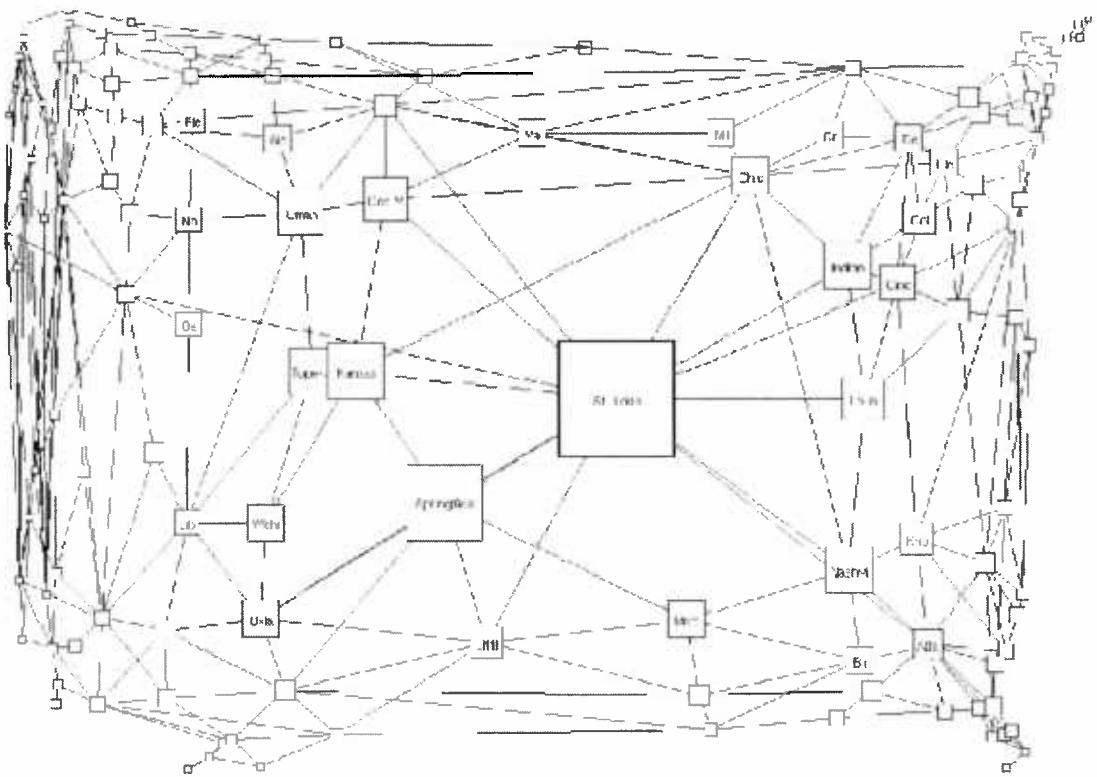
Βελτιωμένη έκδοση της τεχνικής Fisheye View (Εικόνα 42) είναι και η τεχνική Hyperbolic Tree [27], η οποία υλοποιείται σε περιβάλλον 2 και 3 διαστάσεων. Τα χαρακτηριστικά της τεχνικής αυτής είναι:

- Χρησιμοποιείται κυρίως για την απεικόνιση ιεραρχικών δομών δεδομένων.
- Το οπτικό αποτέλεσμα βασίζεται στις αρχές της Hyperbolic γεωμετρίας, από την οποία παίρνει και το όνομα της η τεχνική.



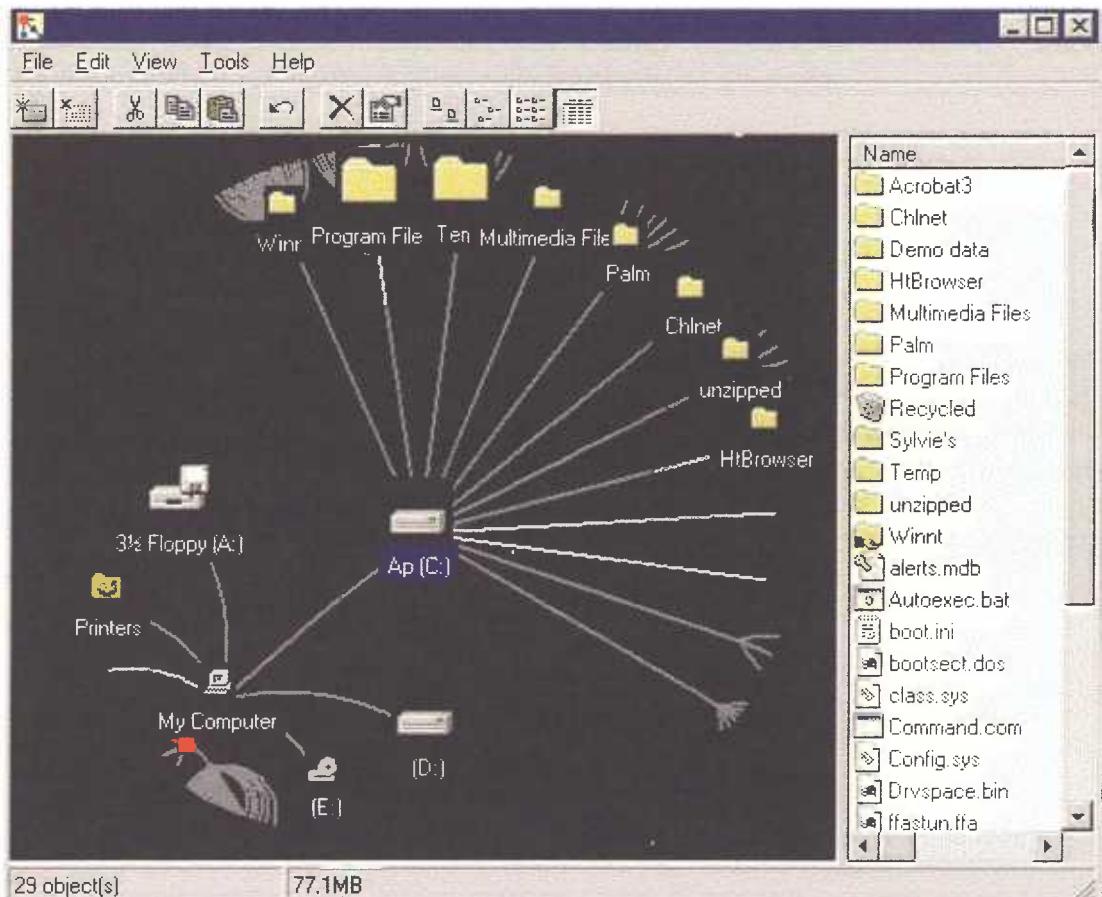


Eukóva 40



Eukóva 41

- Εφαρμόζει τις ιδιότητες της τεχνικής fisheye, π.χ. ένδειξη της σημασίας του κάθε κόμβου μέσω του μεγέθους του.
- Έχει δυνατότητα για απεικόνιση μεγάλου αριθμού δεδομένων, τόσο το δισδιάστατο όσο και το τρισδιάστατο γράφημα.



Εικόνα 42

Το παραπάνω παράδειγμα είναι μία πρωτοποριακή προσπάθεια για χρήση της τεχνικής Hyperbolic Tree προκειμένου να απεικονιστεί η δομή των καταλόγων ενός Η/Υ. Πρόκειται για τον γνωστό εξερευνητή των Windows όπου το αριστερό υπόπαράθυρο δεν απεικονίζει τους καταλόγους με το γνωστό iεραρχικό τρόπο αλλά με την Hyperbolic τεχνική. Τα χαρακτηριστικά που του προσδίδουν το πλεονέκτημα έναντι του κλασικού τρόπου είναι: 1. Η ομορφότερη απεικόνιση των αποτελεσμάτων και 2. Η παροχή της επιπρόσθετης πληροφορίας του φυσικού μεγέθους του κάθε καταλόγου, η οποία παρουσιάζεται από το μέγεθος της απεικόνισής του.

Τεχνική Υπερ-Κύβου (HyperBox)

Το hyperbox είναι μία δισδιάστατη φωτογραφία ενός πολυδιάστατου κύβου [28]. Ένα σύνολο ενωμένων παραλληλογράμμων που αντιστοιχούν σε κάθε διάσταση, συγκεκριμένου μεγέθους και κλίσης, σχηματίζουν αυτόν τον κύβο. Χρησιμοποιείται για αναπαράσταση πολυδιάστατης πληροφορίας.

2.3.8 Δυναμικές Τεχνικές

Το κύριο χαρακτηριστικό της κατηγορίας αυτής είναι η δυνατότητα αλληλεπίδρασης. Στην πραγματικότητα δεν πρόκειται για μια νέα κατηγορία τεχνικών, αλλά κυρίως αφορά υλοποιήσεις των προαναφερθέντων τεχνικών από συστήματα τα οποία επιτρέπουν την αλληλεπίδραση με το χρήστη. Οι τεχνικές αυτές διατηρούν όλα τα χαρακτηριστικά των τεχνικών που υλοποιούν και πολλές φορές συνδυάζουν χαρακτηριστικά περισσότερων από μία τεχνικών, με κύριο στόχο την βελτίωση του παραγόμενου αποτελέσματος.



2.4 Αξιολόγηση και Σύγκριση

Η οπτική απεικόνιση των δεδομένων μεγάλων βάσεων δεδομένων καθώς και των αποτελεσμάτων κάποιων διαδικασιών που εφαρμόζονται σ' αυτά είναι ιδιαίτερης σημασίας, γιατί επιτρέπει στον παρατηρητή την εύκολη και γρήγορη εξερεύνηση και εξαγωγή συμπερασμάτων.

Η ανθρώπινη σκέψη μπορεί πιο εύκολα και πιο γρήγορα να κατανοήσει και να αφομοιώσει αν χρειάζεται μία γραφική οπτική παρουσίαση των δεδομένων, απ' ότι μία γραμμική - σειριακή παράθεση αριθμών. Η εξουκείωση του ανθρώπου με τις εικόνες είναι γνωστή από παλιά και αυτό δηλώνει σαφώς η παροιμία: "Μία εικόνα αξίζει όσο 1000 λέξεις". Βέβαια, ορισμένες τεχνικές προχώρησαν πιο μπροστά από την παροιμία αυτή συνδυάζοντας εικόνα και λεκτικά για να επιτύχουν ακόμη καλύτερη παρουσίαση.

Στη συνέχεια θα πραγματοποιηθεί μία προσπάθεια συγκριτικής παρουσίασης και κατηγοριοποίησης των τεχνικών που περιγράφηκαν, σύμφωνα με τον τομέα εφαρμογής τους. Βέβαια, η σύγκριση αυτή είναι υποκειμενική μιας δεν υπάρχουν αντικειμενικά κριτήρια. Η σύγκριση δεν βασίζεται στους χρόνους απόδοσης ή την αξιοπιστία της τεχνικής αφού δεν αξιολογείται κάποια υλοποίησή της αλλά στη λογική στην οποία βασίζεται καθώς στα χαρακτηριστικά της.

Στον πίνακα 4 έχουν ταξινομηθεί οι τεχνικές ανά κατηγορία, με βάση την παραπάνω κατηγοριοποίηση [2]. Για κάθε μία ορίζεται ο προτεινόμενος τομέας εφαρμογής της. Στην απεικόνιση χρησιμοποιήθηκε μία σήμανση¹ η οποία επεξηγείται στο τέλος της σελίδας.

¹ ✓ : Προτεινόμενος τομές εφαρμογής

? : Πιθανός τομέας εφαρμογής

✗ : Μη ενδεικνυόμενος τομέας εφαρμογής

Κατηγορία	Τεχνική	Τομέας Εφαρμογής					
		Clustering	Association	Hierarchical Data	Data Mining (Huge Data)	Time Series	Categorical Data
Geometric	Scatterplot Matrices	✓	✗	✗	?	✗	
	Landscapes	?	✗	✗	?	✗	
	Projection	✓	✗	✗	✓	✓	
	Hyperslice	✓	✗	✗	✓	?	
	Parallel Coordinates	✓	✗	✗	?	?	
Icon Based	Stick Figure	✓	✓	✗	✓	✗	
	Shape Coding	✓	✗	✗	?	✓	
	Chernoff Faces	✓	✓	✗	✗	✗	
Pixel Oriented	Query Dependent	✓	✗	?	✓	✓	
	Query Independent	✓	✗	?	✓	✓	
Hierarchical	Dimensional Stacking	✓	✗	✓	✓	✗	
	N - Vision	✓	✗	✓	?	✗	
	Treemap	?	✗	✓	?	✗	
	Cone Trees	✗	✗	✓	?	✗	
	InfoCube	?	✗	✓	✓	✗	

Κατηγορία	Τεχνική	Τομέας Εφαρμογής					
		Clustering	Association	Hierarchical Data	Data Mining (Huge Data)	Time Series	Categorical Data
Graph Based	Basic Graphs	✓	✓	✓	✓	✗	
	Hygraph	✓	✓	✓	✓	✗	
	Seenet	✓	✓	✓	✓	✗	
	Narcissus	✓	✓	✓	✓	✗	
Distortion	Perspective Wall	?	✗	✗	✓	✓	✓
	Table Lens	✓	?	✓	✓	✓	✓
	Fisheye View	✓	✓	✓	✓	?	
	Hyperbolic Trees	?	✓	✓	✓	✗	
	HyperBox	?	✗	✗	✗	✗	✗

Πίνακας 4

ΚΕΦΑΛΑΙΟ 3^ο

Σύστημα Εξόρυξης Γνώσης



Κεφάλαιο 3: Σύστημα Εξόρυξης Γνώσης

3.1 Εισαγωγή

Η επιλογή μίας τεχνικής παρουσίασης δεδομένων δεν είναι τετριμμένη διαδικασία. Υπάρχουν πολλοί παράγοντες οι οποίοι θα πρέπει να ληφθούν υπόψη για να γίνει μία σωστή κίνηση.

Ο πιο σημαντικός από αυτούς είναι η υφή των δεδομένων που πρέπει να παρουσιαστούν. Όπως έγινε κατανοητό από το προηγούμενο κεφάλαιο, η κάθε τεχνική απευθύνεται σε συγκεκριμένες δομές δεδομένων για τις οποίες μπορεί να λειτουργήσει σωστά και αποτελεσματικά. Σε άλλες δομές μπορεί να λειτουργήσει ικανοποιητικά και σε άλλες δεν έχει νόημα να εφαρμοστεί. Για παράδειγμα είναι άσκοπο, μία τεχνική που απευθύνεται σε δεδομένα με ακολουθιακή-γραμμική δομή όπως time-series, να χρησιμοποιηθεί σε δεδομένα που έχουν ιεραρχική υπόσταση. Το βήμα αυτό γίνεται ακόμη δυσκολότερο όταν αναφερόμαστε σε δεδομένα που δημιουργούνται από διαδικασίες εξόρυξης γνώσης γιατί η πληροφορία που εμπεριέχεται σ' αυτά είναι πιο πολύπλοκη και χρειάζεται πιο προσεκτικούς χειρισμούς προκειμένου να μην οδηγηθούμε σε λάθος κατανόηση τους, γεγονός που θα σήμαινε αχρήστευση της τεχνικής.

Ένας, ακόμη, σημαντικός παράγοντας είναι οι χρήστες στους οποίους το σύστημα απευθύνεται. Το γνωστικό επίπεδο, ο επαγγελματικός τομέας, το τμήμα εργασίας που ανήκουν προσδιορίζουν ιδιαίτερες ανάγκες, οι οποίες πρέπει να ληφθούν υπόψη σοβαρά. Για παράδειγμα, δεν είναι εφικτό ο απλός χρήστης του οποίου το γνωστικό επίπεδο δεν είναι υψηλό και ο επαγγελματικός τομέας του δεν είναι σχετικός με την επεξεργασία και ανάλυση δεδομένων να κατανόσει την απεικόνιση των pixel oriented τεχνικών.

Τέλος, σημαντικός παράγοντας είναι οι απαιτήσεις των χρηστών. Χρήση τρισδιάστατων τεχνικών και δυνατότητας άμεσης και εύκολης αλληλεπίδρασης με το σύστημα είναι στοιχεία που όλοι οι χρήστες ζητούν. Αυτά θα πρέπει να συγκεραστούν με την αξιοπιστία, την ταχύτητα ανταπόκρισης και τη λειτουργικότητα που ένα τέτοιου είδους σύστημα πρέπει να παρέχει.

Στα επόμενα κεφάλαια θα γίνει αναφορά στο υπάρχον σύστημα εξόρυξης γνώσης, θα παρουσιαστούν οι τεχνικές που έχουν επιλεγεί και θα εξηγηθούν οι λόγοι επιλογής τους.



3.2 Περιγραφή Συστήματος

Το υπάρχον σύστημα έχει ως σκοπό την εφαρμογή τεχνικών clustering, classification και association rules χρησιμοποιώντας ασαφή λογική (fuzzy logic). Ο τρόπος, με τον οποίο λειτουργούν μέχρι σήμερα όλα τα πληροφοριακά συστήματα είναι προσαρμοσμένος στις ανάγκες των μηχανημάτων και όχι του ανθρώπου. Χρειάζεται δηλ. προσπάθεια εκμάθησης του συστήματος για να μπορέσει να γίνει σωστή χρήση του και αυτό οφείλεται στο γεγονός ότι ο υπολογιστής δεν μπορεί να καλύψει το εύρος σκέψης του ανθρώπου. Πολλές έρευνες έχουν γίνει (ξεκινώντας από τη δεκαετία του '70) για την μείωση του κενού της επικοινωνίας των ανθρώπων με τις μηχανές, αρκεί να αναφερθούν όλα τα συστήματα που κατά καιρούς έχουν αναπτυχθεί για προσομοίωση φωνής, αναγνώριση φωνής κ.λ.π. Ένας νέος κλάδος έκανε την εμφάνισή του εκείνη την εποχή ονομαζόμενος τεχνητή νοημοσύνη.

Όσον αφορά τα συστήματα επεξεργασίας, ανάλυσης και εξερεύνησης των βάσεων δεδομένων, έχει αναπτυχθεί λογισμικό το οποίο υποστηρίζει κάποιες γλώσσες προγραμματισμού (όπως SQL) προκειμένου να μπορούν να εξαχθούν συμπεράσματα για τη δομή των δεδομένων και τις τιμές τους. Ωστόσο η τεχνολογία αυτή δεν επιτρέπει στον χρήστη να πάρει πληροφορία την οποία δεν γνωρίζει. Έτσι, πρέπει αρχικά να δημιουργηθούν πιθανά σενάρια και στη συνέχεια να ετοιμαστούν πιθανές ερωτήσεις προς το σύστημα οι οποίες να επιβεβαιώνουν ή να αναιρούν τις ερωτήσεις αυτές.

Οι διαδικασίες Data Mining εμφανίστηκαν για να καλύψουν αυτό το κενό δημιουργώντας την απαραίτητη τεχνολογία αναζήτησης και εντοπισμού γνώσης μέσα από τα δεδομένα. Ο χρήστης δεν είναι πλέον αναγκασμένος να δημιουργεί σενάρια, αλλά απλώς τρέχει κάποιες διαδικασίες Data Mining και το σύστημα δημιουργεί τα σενάρια αυτά, ελέγχει την ορθότητά τους και παρουσιάζει εκείνα που ισχύουν. Και πάλι όμως το σύστημα λειτουργεί ψηφιακά. Δεν είναι δηλαδή σε θέση να καταλάβει ότι μεταξύ των τιμών 1000 και 1000,1 υπάρχει ελάχιστη διαφορά, η οποία στον ανθρώπινο νου δεν σημαίνει τίποτα. Στην ψηφιακή τεχνολογία αυτές οι τιμές αναπαρίστανται από διαφορετικές ψηφιακές λέξεις, άρα είναι διαφορετικές.

Η αδυναμία του συστήματος να κατανοήσει τον συνεχή – αναλογικό τρόπο σκέψης του ανθρώπου είναι σημαντικό μειονέκτημα. Για παράδειγμα, ένας άνθρωπος που είναι 35 χρονών, μπορεί να θεωρηθεί ότι είναι μεσαίας ηλικίας ή ότι είναι σχετικά μικρός. Έτσι, αν επιθυμούσα να θέσω την ακόλουθη ερώτηση στο σύστημα μου: «Φέρε μου όλους τους μικρούς ανθρώπους», θα περίμενα να πάρω σαν αποτέλεσμα και την συγκεκριμένη εγγραφή. Ωστόσο, ο τρόπος που ένα σύστημα δομεί τις απαντήσεις του στηρίζεται σε ακριβείς αριθμούς. Άρα, η παραπάνω ερώτηση, θα είχε νόημα μόνο αν είχα προηγουμένως ορίσει ότι μικροί άνθρωποι είναι αυτοί οι οποίοι ανήκουν στις ηλικίες από 0 μέχρι 30. Και βέβαια μία τέτοια δόμηση σημαίνει ότι μπορώ να χάσω πολύ σημαντική πληροφόρηση γιατί όπως φαίνεται και από το παράδειγμα δεν θα αντλούσα την πληροφορία που μου προσφέρει η εγγραφή του 35χρονου.



Το καλύτερο σ' αυτή την περίπτωση θα ήταν να έχω ένα αποτέλεσμα που να μου λέει ότι βρέθηκαν x άνθρωποι που είναι μικροί και y άνθρωποι που είναι σχετικά μικροί. Σε ένα ακόμη παραπάνω βήμα θα έπρεπε να μου πει κατά πόσο οι άνθρωποι αυτοί ικανοποιούν τον κανόνα «είναι μικροί» ή κατά πόσο ικανοποιούν τον κανόνα «είναι σχετικά μικροί». Αυτό ακριβώς προσπαθεί να κάνει η ασαφής λογική, η οποία νιοθετείται από το σύστημά μας.

Ο βαθμός ικανοποίησης κάποιου κανόνα ονομάζεται Degree Of Belief (d.o.b.) και είναι ένας αριθμός από το 0 έως το 1 (fuzzy domain). Έτσι, ένας άνθρωπος 18 ετών θα λέμε ότι ανήκει στον κανόνα «είναι μικροί» με d.o.b.=0.75, ενώ ένας άνθρωπος 35 ετών θα λέμε ότι ανήκει στον κανόνα «είναι μικροί» με d.o.b.=0.15 ενώ ταυτόχρονα ανήκει στον κανόνα «είναι σχετικά μικροί» με d.o.b.=0.50.

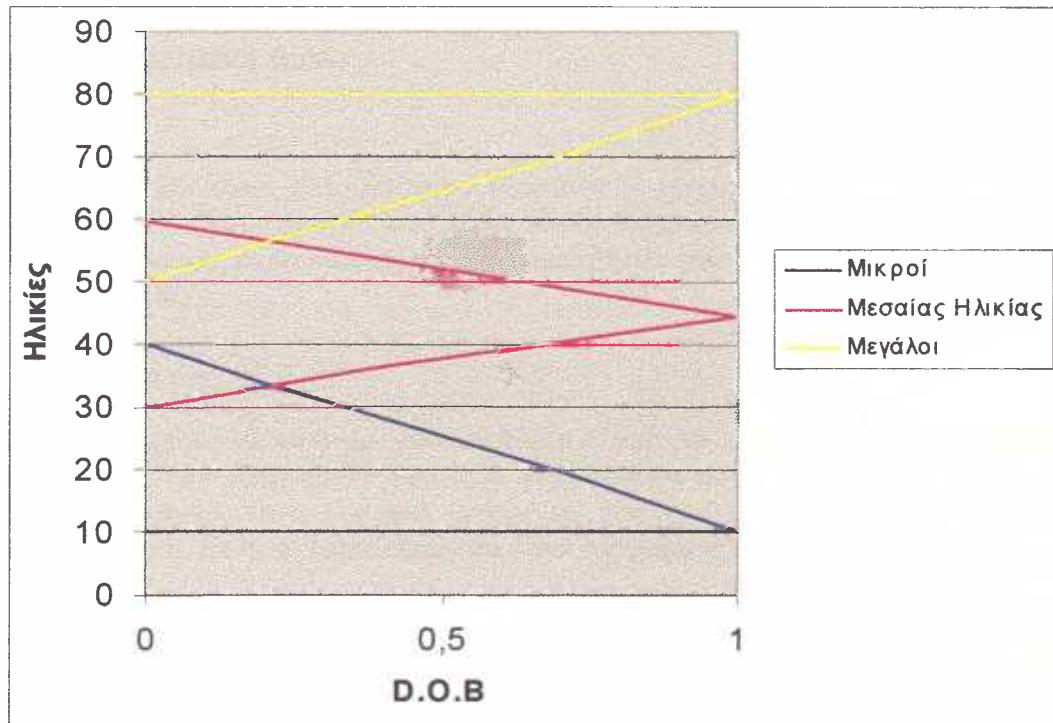
Με αυτόν τον τρόπο αποφεύγεται το να θέτουμε ακριβή όρια στις διαδικασίες κατηγοριοποίησης των δεδομένων που εκτελούνται από τα μοντέλα εξόρυξης γνώσης. Το ακόλουθο παράδειγμα παρουσιάζει ένα πιθανό πλεονέκτημα της μεθόδου αυτής σε σχέση με την κλασική. Έστω, ότι τα δεδομένα δημιουργούν ένα cluster: οι άνθρωποι οι οποίοι ανήκουν στις ηλικίες από 30 μέχρι 35 αγοράζουν σε μεγάλο βαθμό ξηρούς καρπούς και έστω ότι ζητάω από το σύστημα να μου παρουσιάσει τους ανθρώπους που ηλικιακά κυμαίνονται από 0 ως 30 σε σχέση με τα τρόφιμα που αγοράζουν. Το αποτέλεσμα θα αποκρύψει το σημαντικό στοιχείο για τους ανθρώπους από 30 μέχρι 35 και αυτό θα συμβεί επειδή ο θέτων την ερώτηση δεν γνώριζε και δεν μπόρεσε να φανταστεί ότι οι ηλικίες αυτές θα παρουσιάζαν κάποιο cluster σε σχέση με τα τρόφιμα.

Η σωστή ερώτηση που το σύστημα θα έπρεπε να επιτρέπει να τεθεί είναι: «Παρουσίασέ μου τα τρόφιμα που καταναλώνουν μικροί σε ηλικία άνθρωποι». Η απάντηση στηριζόμενη σε ασαφή λογική θα έπρεπε να παρουσιάζει τις ηλικίες ανάλογα με το βάρος τους (d.o.b.).

Το ερώτημα που τίθεται σε αυτό το σημείο είναι το πώς υπολογίζεται το d.o.b. της κάθε τιμής. Για κάθε μεταβλητή, ορίζεται το σύνολο τιμών της. Στη συνέχεια, καθορίζονται οι κατηγορίες με χρήση της συνάρτησης εύρεσης clusters που υλοποιεί το σύστημα. Τέλος ορίζεται μία συνάρτηση μετασχηματισμού για κάθε κατηγορία που δημιουργήθηκε, η οποία έχει σαν σκοπό να προσδιορίζει το d.o.b. Ο πίνακας 5 και η εικόνα 44 απεικονίζουν την κατάσταση αυτή.

	Μικρός	Μεσαίας ηλικίας	Μεγάλος
Min	1	30	50
Max	40	60	80
Συνάρτηση	Decreasing	triangle	Increasing

Πίνακας 5



Εικόνα 43

Μετά τον καθορισμό των κατηγοριών πρέπει να υπολογιστεί ποια από αυτές έχει μεγαλύτερη δύναμη. Δηλαδή, από το σύνολο των δεδομένων θα δούμε ποια κατηγορία είναι αυτή στην οποία ανήκουν οι περισσότερες εγγραφές και με το μεγαλύτερο d.o.b. Ορίζεται η μετρική E (Ενέργεια) η οποία υπολογίζει για κάθε κατηγορία ξεχωριστά τη τιμή αυτή.

Πλέον, το σύστημα είναι σε θέση να απαντήσει στην ερώτηση «ποιος είναι ο βαθμός στον οποίο το σύνολο των δεδομένων μου περιέχει μεγάλους ανθρώπους», παρέχοντας στον χρήστη τη μετρική E (Ενέργεια). Επίσης, είναι δυνατή η δημιουργία συγκρίσεων συνεπώς μπορεί να απαντηθεί η ερώτηση «Βρες την ηλικιακή κατηγορία που συναντάται περισσότερο», συγκρίνοντας τις Ενέργειες κάθε κατηγορίας ηλικιών και επιλέγοντας τη μεγαλύτερη. Με τη χρήση της μετρικής E και της λογικής Boolean επιτυγχάνεται να γίνουν συνδυαστικές ερωτήσεις του τύπου: «βρες αν οι ακριβότερες αγορές γίνονται πρωί ή απόγευμα» (Εακριβές and πρωί or Εακριβές and απόγευμα). Τέλος, με την ίδια λογική μπορούν να ανακαλυφθούν συσχετίσεις (association rules) μεταξύ των κατηγοριών αυτών.

Στο επόμενο κεφάλαιο θα γίνει αναφορά στα ιδιαίτερα χαρακτηριστικά που πρέπει να πλήρη η οπτική παρουσίαση των παραπάνω τεχνικών και θα προταθούν λύσεις.

3.3 Προτεινόμενη Λύση

Στο κεφάλαιο αυτό θα παρουσιαστούν οι τεχνικές που επιλέχθηκαν και θα εξηγηθούν οι λόγοι επιλογής τους. Κάθε αλγόριθμος του υπάρχοντος συστήματος και η τεχνική οπτικής παρουσίασης των αποτελεσμάτων του περιγράφονται σε ξεχωριστά υπό-κεφάλαια.

3.3.1 Clustering

Η πρώτη λειτουργία που εκτελείται από το σύστημα είναι η εύρεση - καθορισμός των κατηγοριών του κάθε γνωρίσματος από το σύνολο δεδομένων. Οι κατηγορίες αυτές είναι αυστηρά ορισμένες, δηλ. δεν λαμβάνουν υπόψη τους την ασαφή λογική. Έτσι, το αποτέλεσμα θα είναι της μορφής:

1. 1 ως 15 χρονών → κατηγορία 1
2. 16 ως 30 χρονών → κατηγορία 2
3. 31 ως 50 χρονών → κατηγορία 3
4. κ.λ.π.

Η ασαφής λογική εφαρμόζεται με βάση κάποιο βαθμό επικάλυψης ο οποίος επιλέγεται από τον χρήστη, αφού σχηματιστούν οι βασικές κατηγορίες. Με βάση το βαθμό επικάλυψης προσδιορίζεται ο βαθμός συμμετοχής (d.o.b.) για κάθε αντικείμενο - εγγραφή της βάσης δεδομένων στην κάθε κατηγορία.

Έτσι, η διαδικασία εύρεσης - ορισμού των κατηγοριών επιστρέφει πληροφορίες για τα κέντρα της κάθε κατηγορίας και για τους βαθμούς συμμετοχής του κάθε στοιχείου δεδομένων στις κατηγορίες αυτές. Αξίζει να σημειωθεί ότι ο προσδιορισμός των κατηγοριών γίνεται μετά από εφαρμογή διαδικασιών κανονικοποίησης των τιμών των δεδομένων, προκειμένου να αντικατοπτρίζεται σωστά η πραγματικότητα, συνεπώς θα πρέπει κατά τη διαδικασία οπτικής παρουσίασης των αποτελεσμάτων να χρησιμοποιηθεί κάποιος τρόπος σωστής κατανομής των τιμών στους άξονες.

Τέλος, οι διαδικασίες αυτές μπορούν να εφαρμοστούν σε δεδομένα μίας, δύο ή περισσότερων διαστάσεων. Δηλαδή, εφαρμόζεται σε ένα γνώρισμα κάποιου πίνακα ή σε περισσότερα του ενός γνωρίσματα. Έτσι, μία πιθανή κατηγορία είναι οι ηλικίες 1 έως 15 όταν αναφερόμαστε μόνο στη διάσταση ηλικίες, ενώ αν χρησιμοποιηθεί και η διάσταση μισθός τότε μία κατηγορία είναι 1 έως 15 και 0 έως 50.000 δρχ.



Συνοψίζοντας τα παραπάνω ορίζουμε τα ακόλουθα χαρακτηριστικά για τις διαδικασίες κατηγοριοποίησης:

- ✓ Δυνατότητα προσδιορισμού των διαστάσεων στις οποίες θα γίνει κατηγοριοποίηση. Συνεπώς, δύναται τα αποτελέσματα να έχουν πληροφορίες για μία, δύο ή περισσότερες διαστάσεις.
- ✓ Πάντα επιστρέφεται ως πληροφορία ο αριθμός των κατηγοριών και τα κέντρα τους στις k διαστάσεις.
- ✓ Για κάθε εγγραφή είναι γνωστή η κατηγορία στην οποία ανήκει.
- ✓ Με τη βοήθεια των νόμων της ασαφούς λογικής μπορεί να παραχθεί ο βαθμός συμμετοχής της κάθε εγγραφής στις ορισμένες κατηγορίες.
- ✓ Χρειάζεται κανονικοποίηση των αξόνων για την σωστή αναπαράσταση των κατηγοριών (σε 2 ή περισσότερες διαστάσεις).

Η δυνατότητα ορισμού πολλών διαστάσεων για τη διαδικασία κατηγοριοποίησης των δεδομένων δημιουργεί μία πολυπλοκότητα στην αναπαράστασή τους δεδομένου ότι θα ήταν καλό να χρησιμοποιηθεί μία τεχνική, η οποία θα έχει τη δυνατότητα να προσαρμόζεται στις εκάστοτε επιλεγμένες διαστάσεις. Τέτοιου είδους τεχνικές, σύμφωνα με το προηγούμενο κεφάλαιο, είναι οι pixel oriented, scatter plot matrices, Hyperslice, stick figures, dimensional stacking και οι table lens.

1 Διάσταση

Οι Pixel oriented τεχνικές ενδείκνυνται για παρουσίαση κατηγοριών στα δεδομένα μίας διάστασης. Χάρη στους αλγόριθμους γεμίσματος της οθόνης, Peano Hilbert & Morton, που χρησιμοποιούνται δημιουργείται μία αρκετά καλή και σαφή εικόνα του αποτελέσματος με ευκρινή ένδειξη των κατηγοριών, ειδικά όταν χρησιμοποιούνται χρώματα. Πιθανή ύπαρξη ιεραρχικής πληροφόρησης θα μπορούσε να αναπαρασταθεί με χρήση των τεχνικών recursive pattern. Τέλος, το χρώμα θα έδειχνε το βαθμό συμμετοχής του κάθε αντικειμένου στην κάθε κατηγορία.

Ωστόσο, η τεχνική αυτή έχει μερικά μειονεκτήματα, όπως το γεγονός ότι δημιουργείται από δεδομένα για τα οποία δεν γνωρίζουμε τις κατηγορίες. Στην δική μας περίπτωση, οι κατηγορίες έχουν βρεθεί και χρειάζεται να αναπαρασταθούν. Για το λόγο αυτό χρειάζεται να τροποποιηθεί λίγο η λογική της τεχνικής αυτής. Μία πιθανή λύση είναι η ακόλουθη:

1. Να δημιουργείται ένα τμήμα στην οθόνη για κάθε ορισμένη κατηγορία. Έτσι αν είχαμε k κατηγορίες θα έπρεπε να δημιουργηθούν k υπό-παράθυρα.
2. Κάθε υπό-παράθυρο να γεμίζει με τα δεδομένα της κάθε κατηγορία ακολουθώντας την ‘Spiral τεχνική’ (Εικόνα 28). Τα δεδομένα με τον μεγαλύτερο βαθμό συμμετοχής θα πρέπει να τοποθετούνται στο κέντρο του υπό-παραθύρου που αναπαριστά την εκάστοτε κατηγορία.



3. Το χρώμα των pixels θα χρησιμοποιηθεί για την αναπαράσταση των τιμών του γνωρίσματος που κατηγοριοποιείται. Σε περίπτωση που τα δεδομένα που κατηγοριοπούνται δεν έχουν συνεχή χροιά τότε το χρώμα μπορεί να χρησιμοποιηθεί για την αναπαράσταση κάποιας εξαρτημένης μεταβλητής.
4. Ο τρόπος κατανομής των υπό-παραθύρων στην οθόνη θα σχετίζεται με τις αποστάσεις των κατηγοριών μεταξύ τους. Έτσι, οι κατηγορίες που απέχουν μεγαλύτερη απόσταση η μία από την άλλη θα πρέπει να αντιστοιχίζονται σε υπό-παραθύρα που επίσης απέχουν πολύ μεταξύ τους. Με τον τρόπο αυτό δημιουργείται μία οπτική εικόνα υπέρ-κατηγοριών. Αν για παράδειγμα τα περισσότερα δεδομένα της εφαρμογής περιείχαν ηλικίες μεταξύ 10 και 30, και ταυτόχρονα οι ηλικίες αυτές δημιουργούσαν 3 κατηγορίες, το οπτικό αποτέλεσμα θα έδινε μία εικόνα ομαδοποίησης των κατηγοριών αυτών.
5. Τέλος, η πυκνότητα και το μέγεθος των παραθύρων, θα έδινε μία επιπρόσθετη εικόνα κατανομής των δεδομένων στις κατηγορίες. Το τελικό αποτέλεσμα θα είναι μία απεικόνιση μικρών γαλαξιών στην οθόνη ο καθένας από τους οποίους θα αναπαριστά μία κατηγορία.

2 ή 3 Διαστάσεις

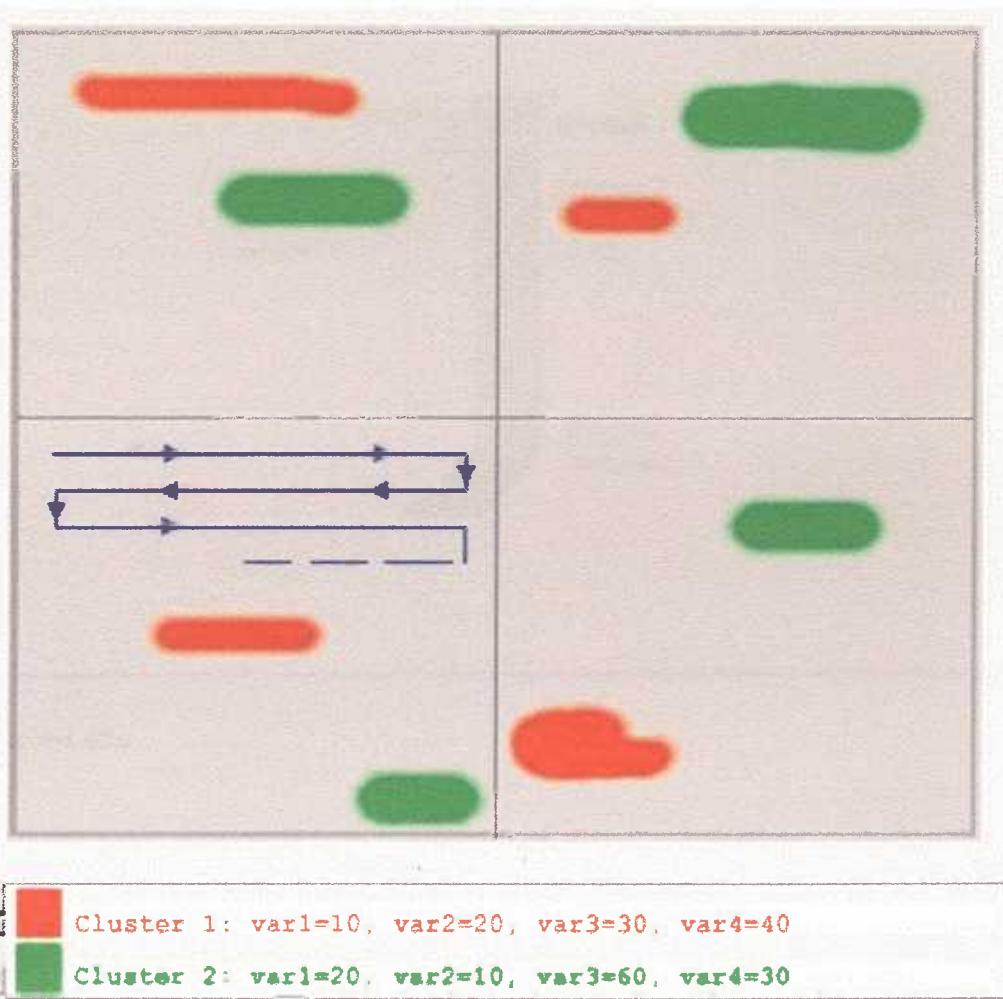
Η παραπάνω λύση ικανοποιεί την περίπτωση που έχουμε μία μόνο διάσταση. Η τεχνική των scatter plot matrices και των Hyperslices δείχνουν να ικανοποιούν την περίπτωση περισσοτέρων διαστάσεων. Τα scatter plot matrices είναι ιδανικά για αναπαράσταση κατηγοριών σε δύο διαστάσεις και κυρίως όταν οι κατηγορίες αυτές έχουν δημιουργηθεί ανά ζεύγη γνωρισμάτων, π.χ. ύψος και ηλικία, ηλικία και μισθός, μισθός και ύψος κ.λ.π. Αδυνατούν όμως να αναπαραστήσουν περισσότερα επίπεδα σε μία εικόνα, δηλ. αν ορίζεται μία κατηγορία για τις ηλικίες 10-13 και τους μισθούς 100 με 110 χιλιάδες και το ύψος 1,80 (τρεις διαστάσεις) τότε δεν υπάρχει δυνατότητα απεικόνισης της πληροφορίας. Από την άλλη η τεχνική των Hyperslices απεικονίζει φέτες του πολυδιάστατου χώρου, δηλ. στην ουσία παίρνει προβολές της κάθε φέτας στον δισδιάστατο χώρο. Συνεπώς και σ' αυτή την περίπτωση δεν υπάρχει δυνατότητα αναπαράστασης πολλών επιπέδων ταυτόχρονα.

Η λογική που εφαρμόζεται στην τεχνική Dimensional Stacking είναι πιο ευέλικτη για την παρουσίαση περισσοτέρων των 2 διαστάσεων. Ωστόσο εμπεριέχει ιεραρχική δομή παρουσίασης μιας και αποτελείται από ενθυλακωμένα δισδιάστατα διαγράμματα.

Από τα παραπάνω, καταλήγουμε στο συμπέρασμα ότι είναι η ταυτόχρονη παρουσίαση περισσοτέρων των τριών διαστάσεων σε ένα γράφημα είναι αδύνατη με τη χρήση των προαναφερθέντων τεχνικών. Άλλωστε είναι δύσκολο και για την ανθρώπινη σκέψη να παρατηρήσει και να συλλάβει την έννοια του πολυδιάστατου χώρου. Έτσι, το καλύτερο γραφικό αποτέλεσμα στις 3 διαστάσεις θα ήταν ένα τρισδιάστατο γράφημα όπου οι κατηγορίες θα παρουσιάζονται με μπάλες σε ένα σύστημα τριών αξόνων, με χρήση τεχνολογίας τρισδιάστατων γραφικών. Για περισσότερες από τρεις διαστάσεις προτείνεται η χρήση της τεχνικής Dimensional Stacking, για την περίπτωση που έχουμε ιεραρχική δομή στα δεδομένα.

K Διαστάσεις

Τέλος, για την περίπτωση όπου δεν υπάρχει ιεραρχική δομή στα δεδομένα και θέλουμε να αναπαραστήσουμε περισσότερες των τριών διαστάσεων τότε προτείνεται η χρήση pixel oriented τεχνικών. Συγκεκριμένα, για κάθε διάσταση-γνώρισμα του πίνακα που αναπαρίσταται θα χρησιμοποιείται ένα παράθυρο της οθόνης. Η κάθε κατηγορία θα αντιστοιχίζεται με ένα χρώμα, το οποίο θα παρουσιάζεται στον χρήστη με χρήση μίας μπάρας σε κάποιο σημείο της οθόνης. Εκεί, θα αναγράφονται και οι τιμές ή το εύρος των τιμών για τις οποίες σχηματίζεται η κάθε κατηγορία. Η τεχνική γεμίσματος των παραθύρων θα είναι η τεχνική γεμίσματος από αριστερά προς τα δεξιά και ανάστροφα. Μεγάλο πλεονέκτημα αυτής της τεχνικής είναι ότι μπορούν να παρουσιαστούν και κατηγορίες στις οποίες συμμετέχουν λιγότερα των k γνωρισμάτων του πίνακα. Για παράδειγμα, μπορούν να αναπαρασταθούν κατηγορίες

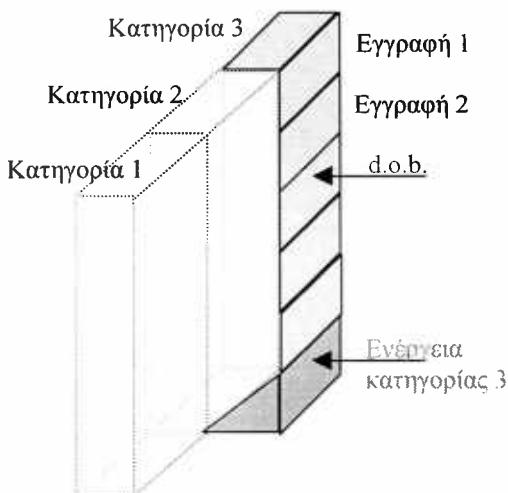


Εικόνα 44

στις οποίες συμμετέχουν δύο μεταβλητές, αλλά και κατηγορίες στις οποίες συμμετέχουν $k+2$ μεταβλητές. Η αναμενόμενη οπτική απεικόνιση των κατηγοριών με χρήση αυτής της τεχνικής παρουσιάζεται στην εικόνα 44.

3.3.2 Classification

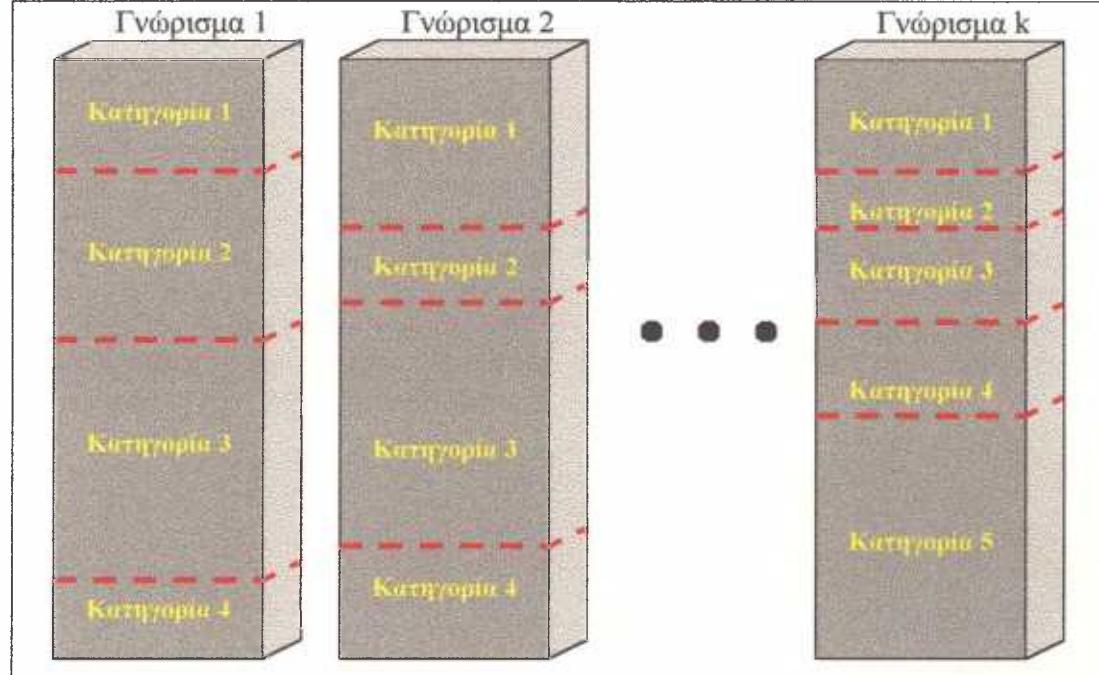
Η διαδικασία classification είναι άμεσα συσχετισμένη με το clustering. Συγκεκριμένα, αποτελεί συνέχεια των αποτελεσμάτων που βγάζει η κατηγοριοποίηση των δεδομένων των τεχνικών clustering. Αναλαμβάνει να υπολογίσει την ενέργεια που έχει κάθε κατηγορία, και να εξάγει συμπεράσματα για την δύναμη τους ως προς το σύνολο των δεδομένων.



Εικόνα 45α

Η εικόνα 45α παρουσιάζει με σαφήνεια την έννοια της ενέργειας. Η εικόνα αυτή πρέπει να παρουσιαστεί στο χρήστη με πιο όμορφο και κατανοητό τρόπο, με χρήση κάποιων τεχνικών ή συνδυασμών τους.

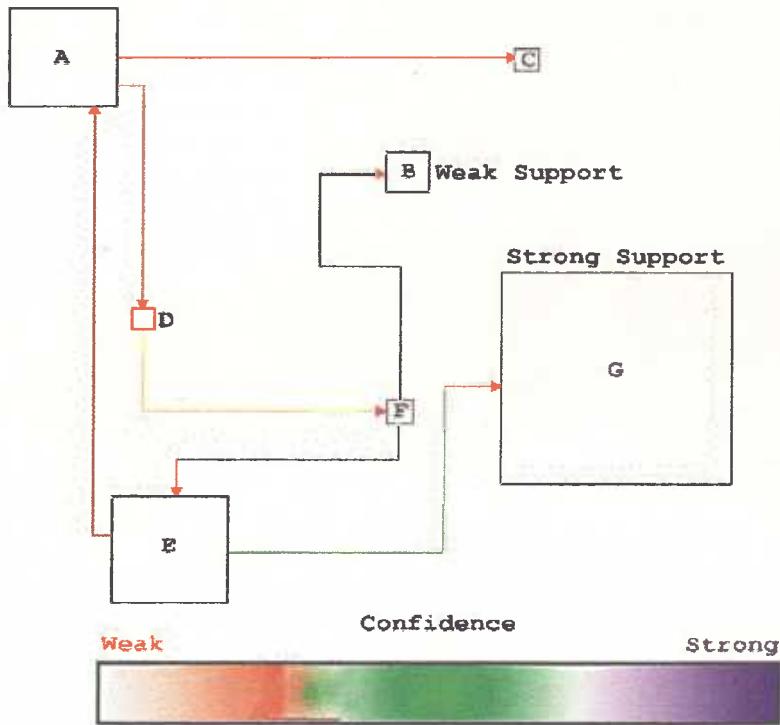
Μία τεχνική που μπορεί να χρησιμοποιηθεί σ' αυτήν την περίπτωση είναι η Treemap. Η τεχνική αυτή χρησιμοποιείται κυρίως για ιεραρχικά δεδομένα, αλλά αν θεωρήσουμε ότι υπάρχουν 2 επίπεδα ιεραρχίας: Γνώρισμα → Κατηγορία, τότε η χρήσή της είναι δυνατή και οδηγεί σε γρήγορες αποφάσεις και συμπεράσματα. Μία οπτική βελτίωση της μεθόδου είναι η χρήση χρωμάτων για την αναπαράσταση των κατηγοριών καθώς και τρισδιάστατα εφέ. Μία προτεινόμενη μορφή της τεχνικής αυτής φαίνεται στην εικόνα 45β.



Εικόνα 45β

3.3.3 Association Rules

Για την οπτική παρουσίαση των κανόνων συσχέτισης καλύτερη τεχνική είναι τα γραφήματα. Το σύστημα που έχει υλοποιηθεί παρέχει απλούς κανόνες 1-1, δηλ. Στοιχείο A → Στοιχείο B, το οποίο σημαίνει ότι το στοιχείο A παρουσιάζει κάποια συσχέτιση με το στοιχείο B, με κάποιο confidence και κάποιο support. Η μετρική Support δείχνει το ποσοστό εμφάνισης επί το σύνολο των δεδομένων του στοιχείου A, ενώ η μετρική confidence δείχνει το ποσοστό των εγγραφών που παρουσιάζουν τη συσχέτιση A → B από το σύνολο των εγγραφών όπου υπάρχει το Στοιχείο A. Οι μετρικές αυτές παρέχονται για κάθε κανόνα συσχέτισης και στην ουσία παρουσιάζουν τον βαθμό εμπιστοσύνης.

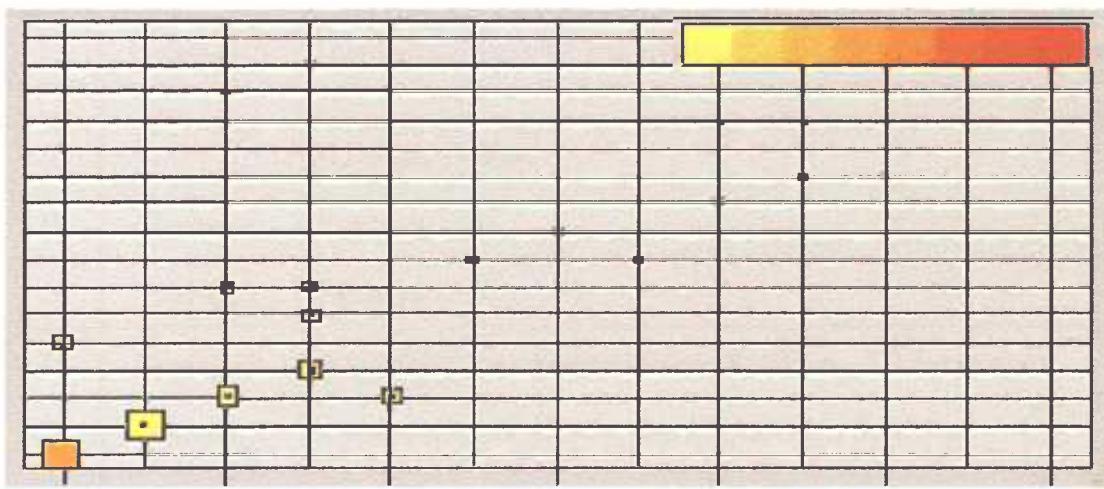


Εικόνα 46

Για την οπτική απεικόνιση τέτοιων σχέσεων μπορεί να χρησιμοποιηθεί οποιαδήποτε από τις προαναφερθείσες τεχνικές γραφημάτων. Τα στοιχεία τα οποία εξετάζονται θα αναπαρίστανται με κόμβους τους γραφήματος και οι σχέσεις τους με ακμές που τα συνδέουν. Επίσης, οι ακμές θα πρέπει να έχουν κατεύθυνση άρα μιλάμε για κατευθυνόμενο γράφημα, χωρίς αυτό να εμποδίζει τη δημιουργία κυκλικών διαδρομών. Το Support και το Confidence πρέπει επίσης να παρουσιάζονται στο

γράφημα. Το support προτείνεται να παρουσιάζεται από το μέγεθος του κόμβου, μιας και αυτό θα αντικατοπτρίζει τη συχνότητα εμφάνισής του στα δεδομένα και το confidence προτείνεται να παρουσιάζεται από το χρώμα της ακμής που συνδέει τους κόμβους (Εικόνα 46).

Μία ακόμη προτεινόμενη απεικόνιση των κανόνων συσχέτισης είναι ένα διάγραμμα 2 αξόνων, όπου στον ένα άξονα παρουσιάζονται τα στοιχεία που βρίσκονται στο αριστερό μέρος της συσχέτισης και στον άλλο άξονα τα στοιχεία που βρίσκονται στο δεξιό μέρος της συσχέτισης. Κάθε συσχέτιση παρουσιάζεται με ένα σημείο στο διάγραμμα, όπου το μέγεθός του δηλώνει το support και το χρώμα του δηλώνει το confidence. Η τεχνική αυτή παρουσιάζεται στην εικόνα 47.



Εικόνα 47



Πίνακας Εικόνων

Σελ.	Σχόλ.	Περιγραφή
<i>Κεφάλαιο 1</i>		
5		Data Mining Process (Περιοδικό Byte Oct 95)
11		Data Mining & Data WareHouse
<i>Κεφάλαιο 2</i>		
13	Γράφημα 1	Παράδειγμα οπτικής παρουσίασης που δείχνει λάθη απεικόνισης
16	Πίνακας	Κατηγορίες τεχνικών οπτικής απεικόνισης των διαδικασιών εξόρυξης γνώσης
19	Εικόνα 1	FastMap – 2D
20	Εικόνα 2	FastMap – 3D
21	Εικόνα 3	Αποτελέσματα από παράδειγμα FastMap - 3D διάγραμμα
21	Πίνακας 1	Αποστάσεις δεδομένων, από Παράδειγμα FastMap
21	Πίνακας 2	Πίνακας αποτελεσμάτων από Παράδειγμα FastMap
22	Πίνακας 3	Pivot lines από παράδειγμα FastMap
23	Εικόνα 4	Projection Views
24	Εικόνα 5	Parallel Coordinates
25	Εικόνα 6	Scatterplot Matrices
25	Εικόνα 7	Landscapes
26	Εικόνα 8-9	Hyperslice
27	Εικόνα 10	Chernoff Faces
28	Πίνακας	Πιθανές διαστάσεις για τα Chernoff faces
29	Εικόνα 11	Stick Figures
29	Εικόνα 12	Shape Coding
30	Εικόνα 13	TileBar
31	Εικόνα 14	Οθόνη από εφαρμογή που χρησιμοποιεί την τεχνική TileBars

32	Εικόνα 15	Dimensional Stacking
33	Εικόνα 16	Οθόνη με αποτελέσματα της τεχνικής Dimensional Stacking
34	Εικόνα 17	Worlds within Worlds
35	Εικόνα 18	Treemap
37	Εικόνα 19	Top-Down Treemap
38	Εικόνα 20	Διάγραμμα Venn
38	Εικόνα 21	Cone Tree
40	Εικόνα 22	InfoCube
41	Εικόνα 23	Pixel Oriented – Γενική Ιδέα
42	Εικόνα 24	Screen Filling τεχνικές – Παρουσίαση αποτελεσμάτων αλγορίθμων (Line by Line, Column by Column, Peano Hilbert, Morton)
43	Εικόνα 25	Οθόνη με αποτελέσματα των Screen Filling Τεχνικών
44	Εικόνα 26	Recursive Pattern
45	Εικόνα 27	Οθόνη με αποτελέσματα της τεχνικής Recursive Pattern
46	Εικόνα 28	Spiral & Snake Spiral τεχνικές
47	Εικόνα 29	Spiral & Snake Spiral τεχνικές σε δύο διαστάσεις
47	Εικόνα 30	Grouping τεχνικές
50	Εικόνα 31	Orthogonal graph
50	Εικόνα 32	Symmetry-optimized graph
50	Εικόνα 33	Cluster-optimized graph
50	Εικόνα 34	Acyclic graph
51	Εικόνα 35	Hygraph
52	Εικόνα 36	SeeNet
53	Εικόνα 37	Narcissus
55	Εικόνα 38	Perspective Wall
56	Εικόνα 39	Table lens



59	Εικόνα 40-41	NON and Fisheye view
60	Εικόνα 42	Hyperbolic tree για εξερεύνηση συστήματος αρχείων
63-64	Πίνακας 4	Συγκριτική παρουσίαση των τεχνικών οπτικής απεικόνισης μεγάλων βάσεων δεδομένων

Κεφάλαιο 3

67	Πίνακας 5	Παράδειγμα κατηγοριών σε δεδομένα
68	Εικόνα 43	Παράδειγμα απεικόνισης κατηγοριών στα δεδομένα
72	Εικόνα 44	Προτεινόμενος τρόπος απεικόνισης κατηγοριών σε k διαστάσεις
73	Εικόνα 45α	Απεικόνιση την μετρικής Ενέργεια
74	Εικόνα 45β	Προτεινόμενη απεικόνιση των αποτελεσμάτων του αλγόριθμου classification
75	Εικόνα 46	Προτεινόμενη απεικόνιση των αποτελεσμάτων των διαδικασιών εύρεσης συσχετίσεων
76	Εικόνα 47	Προτεινόμενη απεικόνιση των αποτελεσμάτων των διαδικασιών εύρεσης συσχετίσεων



Αναφορές

- [1] Daniel A. Keim, *Visual Techniques for Exploring Databases*, KDD 97, Tutorial Notes.
- [2] Daniel A. Keim, Hans-Peter Kriegel, *Visualization Techniques for Mining Large Databases: A comparison*, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No 6, Dec 1996
- [3] George Robertson, Stuart K. Card, Jock D. Mackinlay, *Information visualization using 3D interactive animation*, Communications of the ACM/April 1993/ Vol.36, No4, pp 57-71
- [4] Diethelm Irony Ostry, thesis, *Some 3 dimensional graph drawing algorithms*, Department of computer sciences, university of Newcastle
- [5] Jarke J. van Wijk, *Hyperslice: Visualization of scalar functions of many variables*, Proc visualization 93, San Jose, CA, 1993, pp119-125
- [6] Rekimoto J., Green M., *The information cube: Using transparency in 3D information visualization*, Proc. 3rd annual workshop on information technologies & systems (WITS 93), 1993, pp 125-132
- [7] Brian Johnson, Ben Shneiderman, *Treemaps: a space – filling approach to the visualization of hierarchical information structures*, Proc. Of the 2nd international IEEE Visualization Conference (San Diego, Oct 1991), pp 284-291
- [8] David Turo, Brian Johnson, *Improving the visualization of hierarchies with Treemaps: Design issues and experimentation*, Proc. of the IEEE conference on visualization, October 1992.
- [9] Clifford Besheres, Steven Feiner, *Auto visual: Rule Based Design of interactive multivariate visualizations*, IEEE computer graphics and applications, Vol. 13, No. 4, 1993, pp 44-58
- [10] Feiner S., Besheres C., *Visualizing n-Dimensional virtual Worlds within World*, Computer Graphics, Vol. 24, No. 2, 1990, pp 76-83
- [11] Clifford Besheres, Steven Feiner, *n-Vision and Auto visual*, www.cs.columbia.edu/graphics/projects/AutoVisual/AutoVisual.html
- [12] LeBlanc J. Ward M.O., Wittels N. *Dimensional Stacking: Exploring N-Dimensional Databases*, visualization 90, San Fracisco, CA, 1990, pp 230-239
- [13] *Dimensional Stacking*, www.mm.uni-paderborn.de/0x83ea6001_0x0001c042



- [14] Marti A. Hearst, *Multi-Paragraph Segmentation of expository text*, Proc. Of the 32nd meeting of the association for computational Linguistics, Los Cruces, NM , June 1994
- [15] Hearst M., *Text Tiling: Segmenting text into multi-paragraph subtopic Passages*, Computational linguistics, 23, pp 33-64, March 1997
- [16] Marti A. Hearst, Christian Plaunt, *Subtopic structuring for full-length document access*, Proceedings of the 16th annual international ACM/SIGR conference, Pittsburgh, PA 1993
- [17] Hearst M., *About Tile Bars*, www.sims.berkeley.edu/~hearst/tb-background.html
- [18] Kurt Thearling, Barry Becker, DeCoste, Mawby, Pilote, Sommerfield, *Visualizing Data Mining Models*
- [19] Daniel Keim, *Pixel oriented visualization techniques for exploring very large Data Bases*, Journal of computational and graphical statistics, 5(1): pp 58-77, 1996
- [20] Daniel Keim, Hans-Peter Kriegel, Micheal Ankerst, *RecursivePattern: A technique for visualizing very large amounts of data*, Proc. Visualization 95, Atlanta, GA, 1995
- [21] Daniel Keim, Hans-Peter Kriegel, *VisDB: Database exploration using multidimensional visualization*, IEEE computer graphics and applications, September 1994, pp.40-49
- [22] Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, Inkeri Verkamo, *Finding interesting rules from large datasets of discovered association rules*, Third international conference on information and knowledge management, Nov. 29 – Dec2, 1994, Maryland
- [23] Christos Faloutsos, King-Ip Lin, *FastMap: a fast algorithm for indexing, Data Mining and visualization of traditional and multimedia datasets*, Journal of computational and graphical statistics for exploring very large databases
- [24] Daniel Keim, Annemarie Herrmann, *The Gridfit Algorithm: an efficient and effective approach to visualizing large amounts of spatial data*, IEEE visualization98
- [25] Pac Chunc Wong, Andrew Crabb, Daniel Bergeron, *Dual multiresolution Hyperslice for multivariate data visualization*, Proc. Of the IEEE information visualization 96, oct 28-oct29, 1996
- [26] Manojit Sarkar and Marc H. Brown, *Graphical Fisheye Views of Graphs*, Proceedings of the ACM/SIGCHI '92 Conference on Human Factors in Computing Systems, May 1992.



- [27] John Lamping, Ramana Rao, and Peter Pirolli, *A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies*, Xerox Palo Alto Research Center, Proc. Human Factors in Computing Systems CHI '95 Conf., Denver, CO, 1995, pp. 401-408.
- [28] Alpern, B. and Carter, L., "The Hyperbox," in Proceedings of Visualization '91, Nielson, G.M. and Rosenblum, L., editors, IEEE Computer Society Press, Los Alamitos, Calif., 1991, pp. 133-139.
- [29] Mackinlay J. D., Robertson G. G., Card S. K.: 'The Perspective Wall: Detail and Context Smoothly Integrated', Proc. Human Factors in Computing Systems CHI '91 Conf., New Orleans, LA, 1991, pp. 173-179.
- [30] Ramana Rao and Stuart K. Card, *Exploring Large Tables with the Table Lens*, Proc. Human Factors in Computing Systems CHI '94 Conf., Boston, MA, 1994, pp. 318-322.
- [31] Hendley R. J., Drew N. S., Wood A. M., Beale R.: 'Narcissus: Visualizing Information', Proc. Int. Symp. On Information Visualization, Atlanta, GA, 1995, pp. 90-94.
- [32] Pickett R. M., Grinstein G. G.: 'Iconographic Displays for Visualizing Multidimensional Data', Proc. IEEE Conf. on Systems, Man and Cybernetics, IEEE Press, Piscataway, NJ, 1988, pp. 514-519.
- [33] Chernoff faces, <http://vlado.fmf.uni-lj.si/VRML/PARIS.97/>

