Visualization of High Dimensional Datasets

Class 10

Challenges of High Dimensional Datasets

High dimensional datasets are common: digital libraries, bioinformatics, simulations, process monitoring, and surveys

Example:

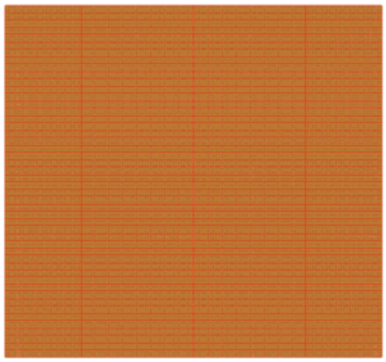
- Ticdata2000 dataset: 86 dimensions
- OHSUMED dataset: 215 dimensions
- SkyServer dataset: 361 dimensions

Challenges of visualizing high dimensional datasets:

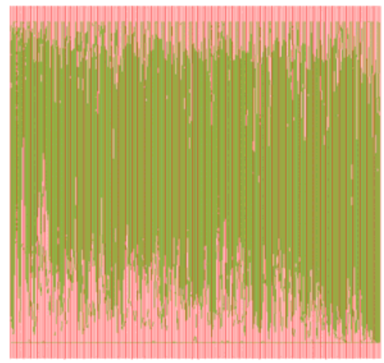
- Clutter on the screen
- Difficult user navigation in the data space

Example

OHSUMED dataset: 215 dimensions



215*215 = 46,225 plots



215 axes

3

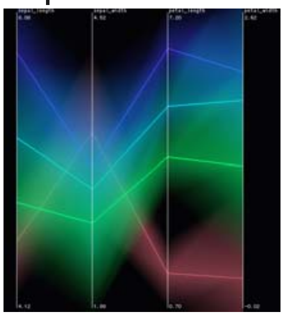
Approach 1: Visual Hierarchical Dimension Reduction (VHDR)

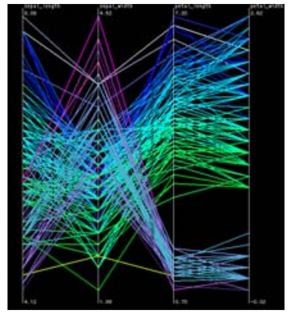
J. Yang, M.O. Ward, E.A. undensteiner and S. Huang

Presented at VisSym'03

4

Inspiration





5

Motivation - Dimension Reduction

Idea:

- Project a high-dimensional dataset to a lower-dimensional subspace
- Visualize data items in the lower-dimensional subspace

Existing Approaches:

- Principal Component Analysis
- Multidimensional Scaling
- Kohonen's Self Organizing Map

Problems:

- Information loss
- No intuitive meaning of generated dimensions
- Little user interaction allowed.

6

Key Ideas of VHDR

- Use dimension hierarchy to convey dimension relationships
- Allow users to learn the dimension hierarchy
- Allow users to select dimensions or dimension clusters to form subspaces of interests

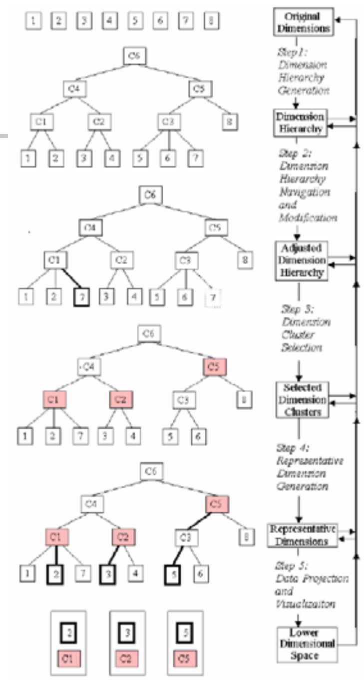
7

VHDR Framework

- Step 1: build dimension hierarchy
- Step 2: navigate and manipulate dimension hierarchy
- Step 3: interactively select clusters from dimension hierarchy to form lower-dimensional subspaces

8

Overview



9

Build Dimension Hierarchy

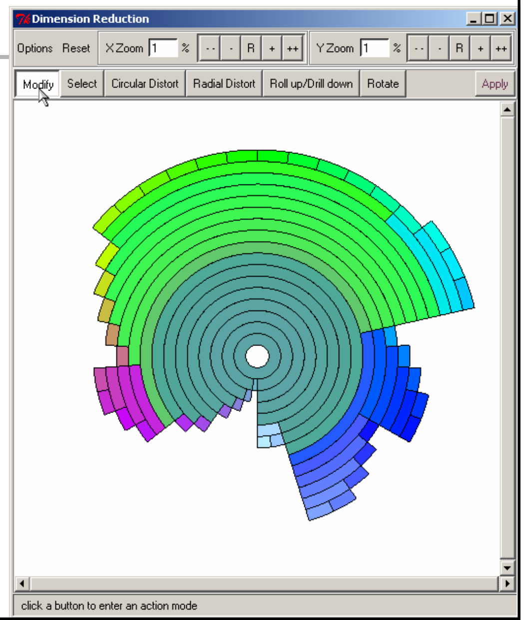
- Automatic dimension clustering
  - Cluster dimensions according to dissimilarities* among them
  - *Dissimilarity - measure of how dimensions are dissimilar to each other
- Manual hierarchy modification
- Discussion:
  - How to calculate dissimilarity between two dimensions?
- Ref:
  - ANKERST, M., BERCHTOLD, S., AND KEIM, D. A. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *InfoVis'98*

10

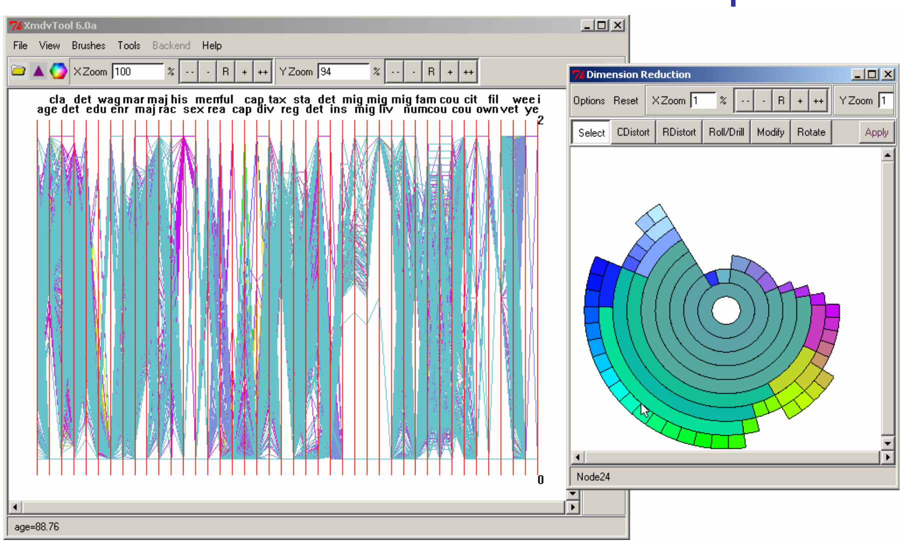
Navigate and Manipulate Dimension Hierarchy

InterRing - Radial space filling hierarchy navigation tool [yang:2002]

- Modification
- Selection
- Radius distortion
- Circular distortion
- Rolling up/Drilling down
- Rotation
- Zooming/Panning



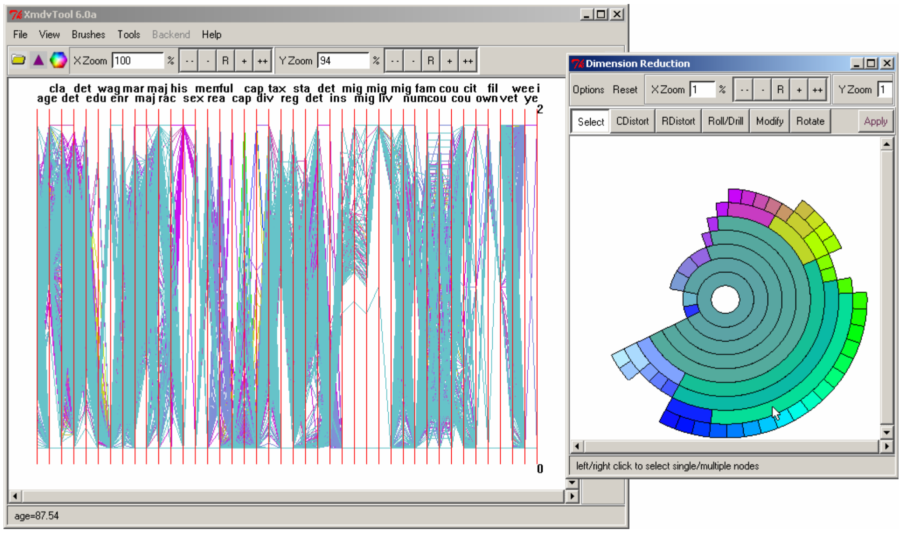
Construct Lower-Dimensional Subspaces



Strategy 1: construct a subspace with closely related dimensions

12

Construct Lower-Dimensional Subspaces



Strategy 2: construct a subspace that covers major variance of the dataset

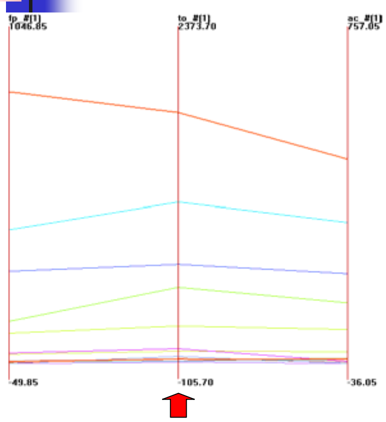
Dimension Cluster Representation

Representative Dimension - a dimension that represents a cluster of dimensions

Approaches to assigning or generating a representative dimension:

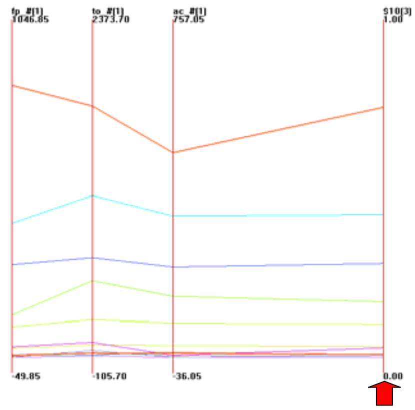
1. Select a dimension from the cluster
2. Average all dimensions in the cluster
3. Use principal component analysis to generate weighted sum of dimensions within a cluster

Examples



representative dimension

Approach - select



representative dimension

Approach - average

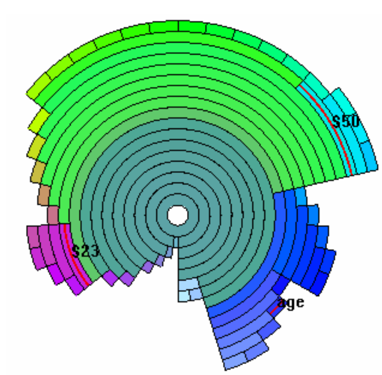
15

Dissimilarity Representation

Goal: visualize dissimilarity of dimensions in a dimension cluster.

Approaches:

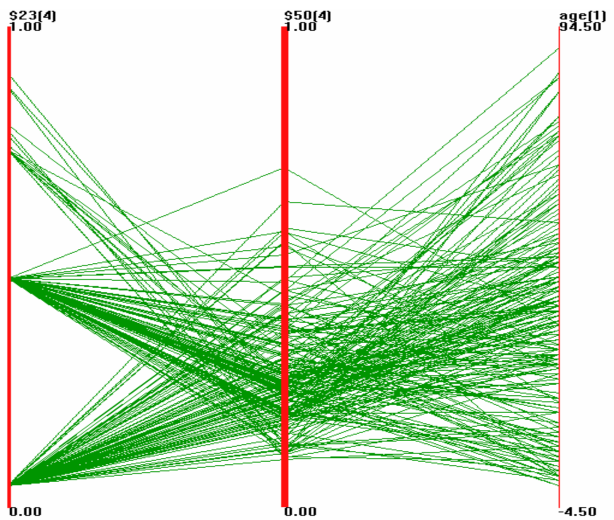
- Axis Width
- Three Axes
- Diagonal Plots
- Outer and Inner Sticks
- Mean-Band



Example: select 3 dimension clusters (dimensions) in Census-Income dataset

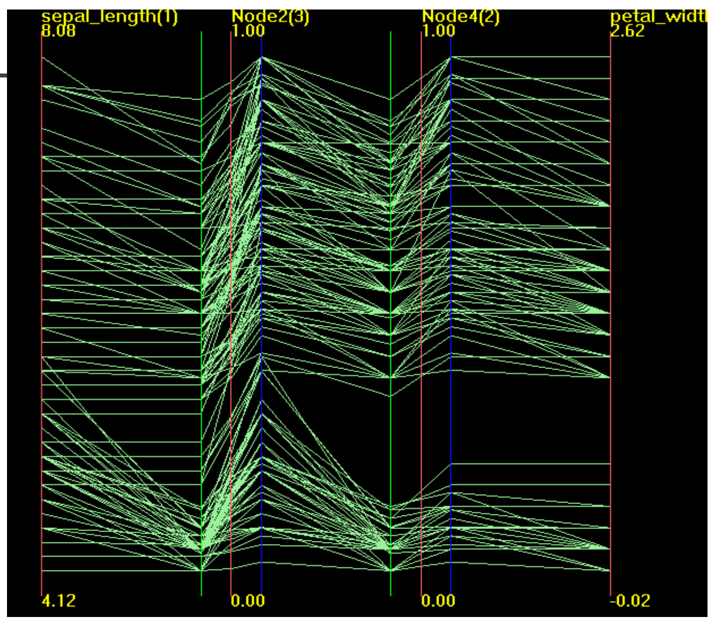
16

Dissimilarity Representation : Axis Width Method



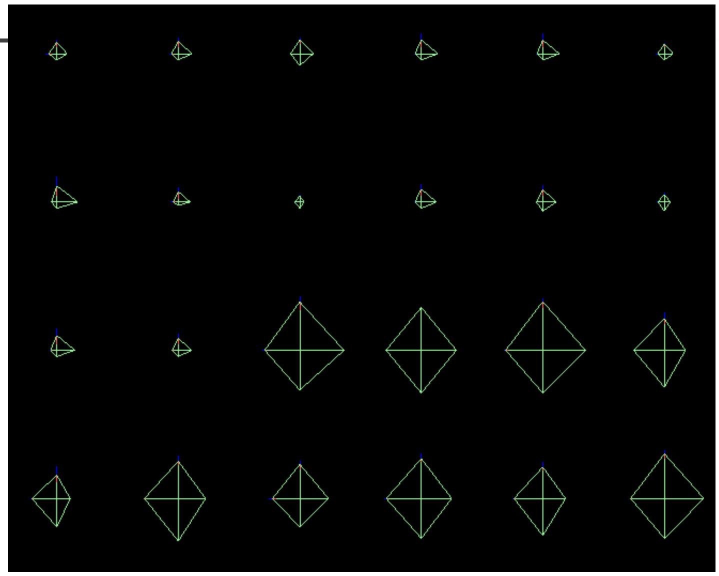
17

Three Axis Method



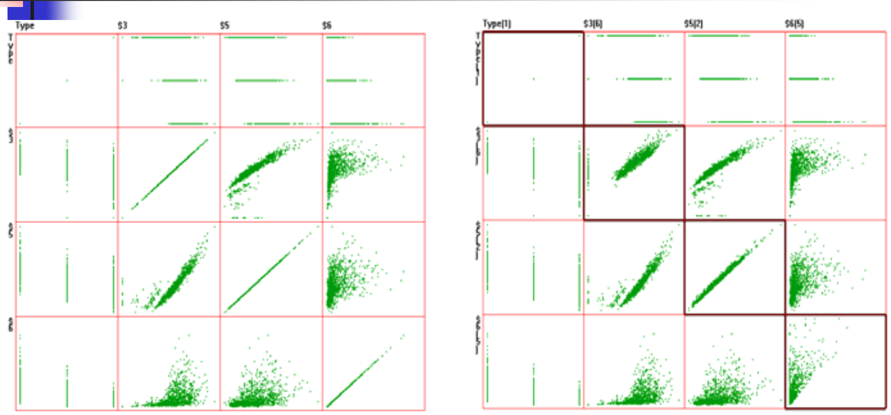
18

Three Axis Method



19

Dissimilarity Representation : Diagonal Plots Method



No dissimilarities representation

Dissimilarities represented in diagonal plots

20

Generality

VHDR is a general framework that can be extended to multiple display techniques

We have applied VHDR to:

- Parallel Coordinates
- Star Glyphs
- Scatterplot Matrices
- Dimensional Stacking

- Hierarchical Parallel Coordinates
- Hierarchical Star Glyphs
- Hierarchical Scatterplot Matrices
- Hierarchical Dimensional Stacking

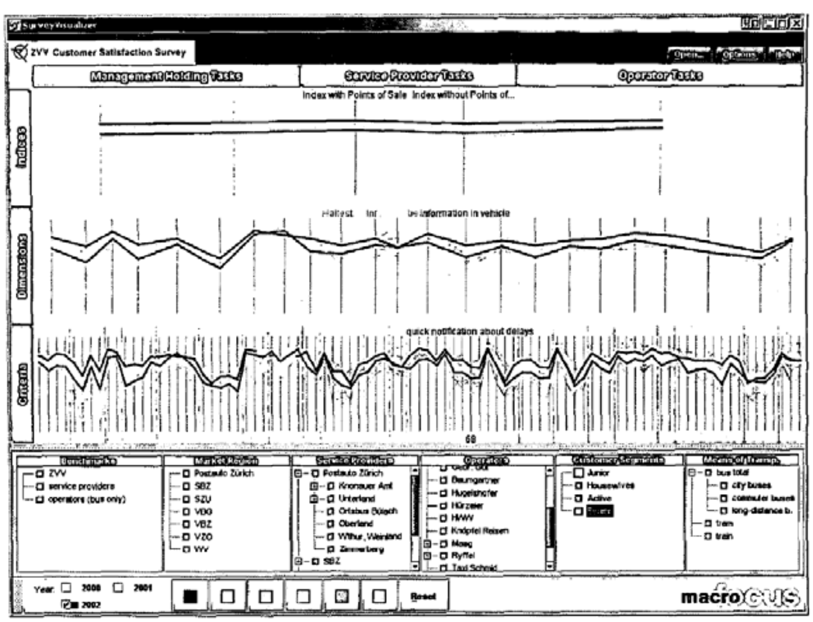
Case studies

- AAUP Dataset: 14 dimensions, 1,131 data items
- Census-Income-Part Dataset: 42 dimensions, 20,000 data items
- Ticdata2000 dataset: 86 dimensions, 5,822 data items
- OHSUMED dataset: 215 dimensions, 298 data items

Other Clustering Approach

- Visualization of Large-Scale Customer Satisfaction Surveys Using a Parallel Coordinate Tree   D. Brodbeck et. al. Infovis 2003

23



24

Approach 2: Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets

Jing Yang, Wei Peng, Matthew O. Ward
and Elke A. Rundensteiner

Presented at InfoVis'03

25

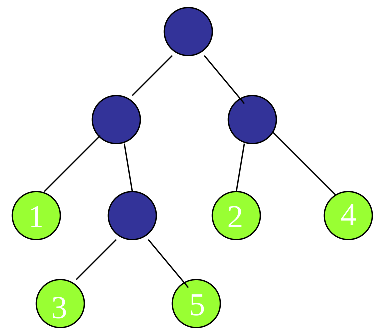
Overview of Our DM Approach

- General: includes dimension ordering, dimension spacing and dimension filtering
- Interactive: allows user interactions throughout the whole process
- Hierarchical: groups dimensions into a hierarchy and builds most algorithms and user interactions upon this hierarchy

26

Dimension Hierarchies

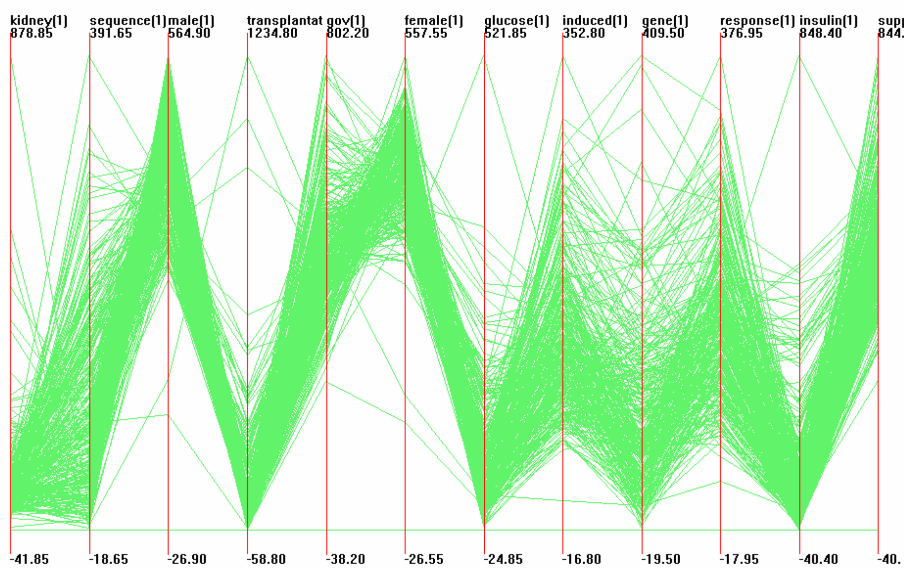
- Dimension hierarchy: similar dimensions form cluster, clusters are grouped into hierarchy
- Dimension hierarchy generation: automatic dimension clustering, manual hierarchy modification



A dimension hierarchy of a 5-dimensional dataset

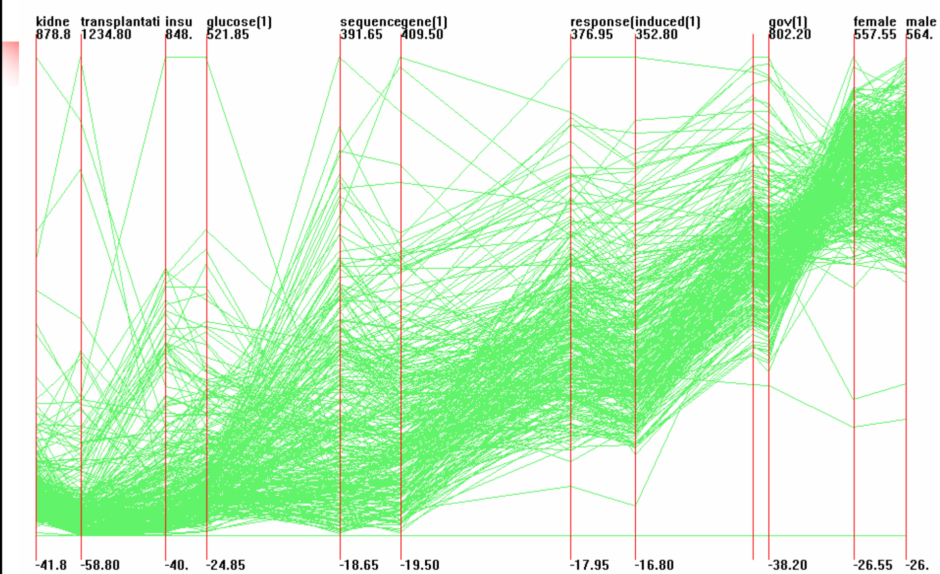
27

Dimension Ordering (1)



A dimension subset of OHSUMED dataset in random order

Dimension Ordering (2)



Same dimensions ordered according to similarity

Dimension Ordering (3)

Order dimensions according to different purposes:

- Similarity-oriented ordering: put similar dimensions close to each other
- Importance-oriented ordering: map more important dimensions to more significant positions or attributes. The order of importance can be decided by Principal Component Analysis (PCA) .

30

Dimension Ordering (4)

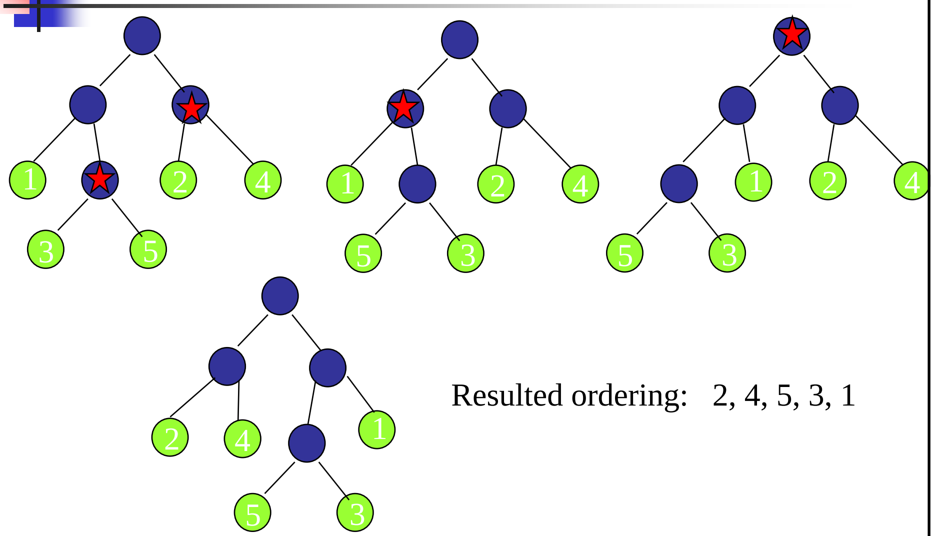
Challenges for ordering high dimensional datasets:

- Similarity-oriented ordering is NP-Complete
- It is hard to decide the order of the importance of a large number of dimensions using PCA

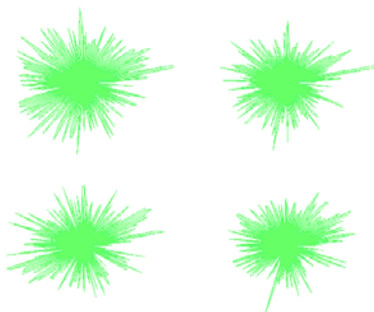
Our solution: reduce the complexity of the ordering problem using the dimension hierarchy

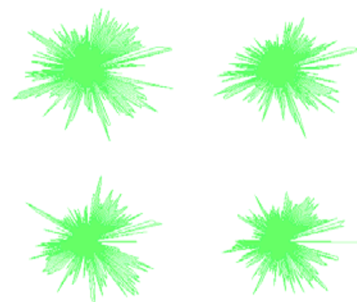
- Order each dimension cluster
- the order of the dimensions is decided in a depth-first traversal of the dimension hierarchy

Hierarchical Ordering Illustration



Dimension Ordering (6)





OHSUMED dataset in random order

OHSUMED dataset in similarity-oriented order

33

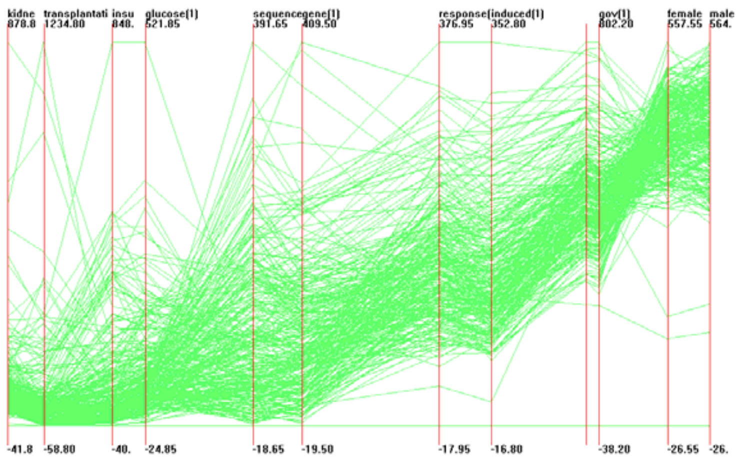
Dimension Spacing (1)

Idea of dimension spacing:

- Convey dimension relationship information by varying the spacing between adjacent axes

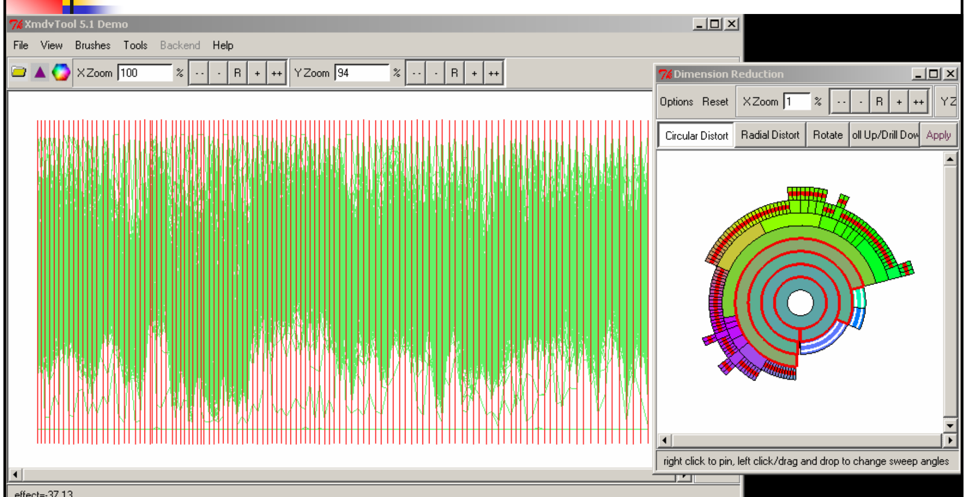
34

Dimension Spacing (2)



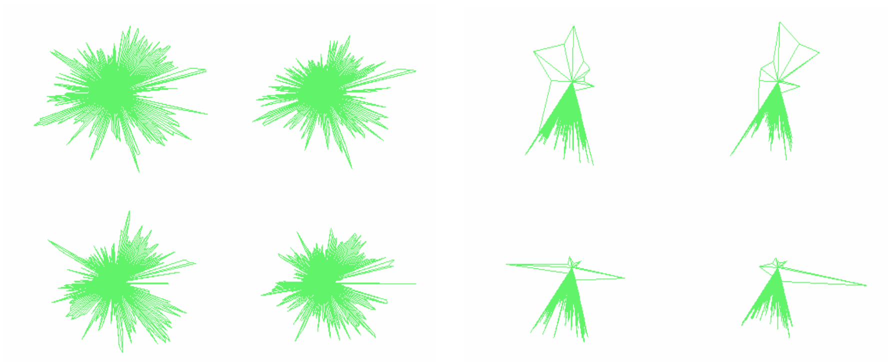
A dimension subset of OHSUMED dataset spaced according to similarity: similar dimensions are close to each other

Structure-Based Space Distortion (1)



Goal: to examine details within context

Structure-Based Space Distortion (3)



OHSUMED dataset Distorted through InterRing

37

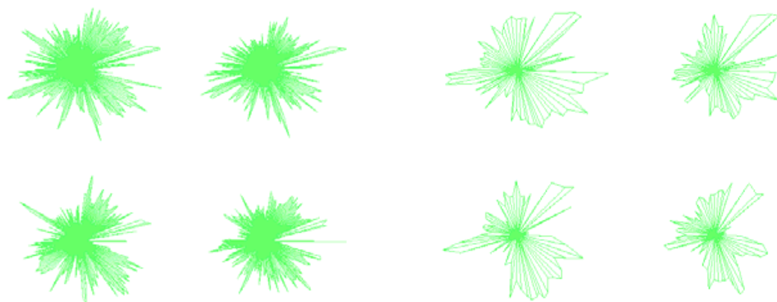
Dimension Filtering (1)

Idea of dimension filtering:

- Similar dimensions can be omitted;
- Unimportant dimensions can be omitted.

38

Dimension Filtering (2)



Unfiltered   Filtered

39

Conclusion

Main contributions of our approach:

- Improves the manageability of dimensions in high dimensional data sets and reduces the complexity of the ordering, spacing and filtering tasks;
- Allows flexible user interactions for dimension ordering, spacing and filtering with dimension hierarchies.

40

Approach 3: Value and Relation (VaR) Display

Jing Yang, Anilkumar Patro, Shiping Huang, Nishant Mehta, Matthew O. Ward and Elke A. Rundensteiner

Presented at InfoVis'04

41

Motivation

Challenges:

- Can high dimensional datasets be visualized without dimension reduction to avoid information loss ?
- Can dimension relationships be visualized in the same display as data values?

42

Challenge - Visualization without Dimension Reduction

Visualize SkyServer dataset (361 dimensions) using existing techniques:

- Parallel Coordinates: 361 axes
- Scatterplot Matrix: 130,321 scatterplots
- Pixel-Oriented techniques *without* overlaps: 50,000 data items: 18,050,000 pixels (23 times of number of pixels in a 1024*768 screen)

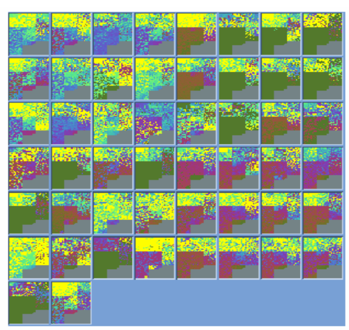
Hint:

- Use Pixel-Oriented techniques and allow overlaps

43

Challenge - Dimension Relationship Visualization

- Sorting dimensions in a 1D or 2D grid [Ankerst 98]
  - Not effective beyond hundreds of dimensions
- Spacing between dimensions [Yang 2003]
  - Only relationships of adjacent dimensions are revealed



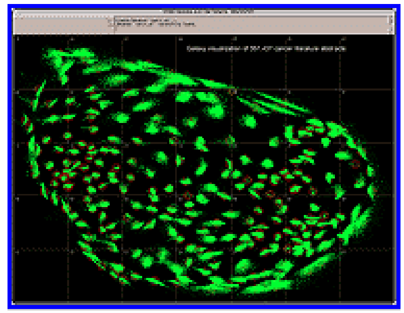
Pixel-Oriented: Sort 50 dimensions in a 2D grid [Ankerst 98]

44

Challenge - Dimension Relationship Visualization (con.)

Recall data item relationship visualization:

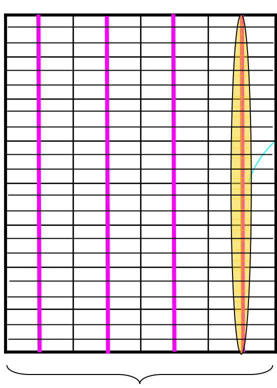
- MDS: SPIRE Galaxies [Wise:95]

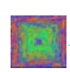


SPIRE Galaxies: Map data items to a 2D display using MDS [Wise: 95]

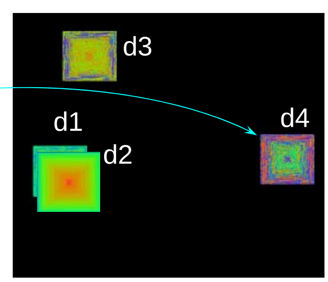
> Hint:
> - Using MDS to layout dimensions

Our Proposal: Value and Relation (VaR) Display





Pixel-Oriented glyph



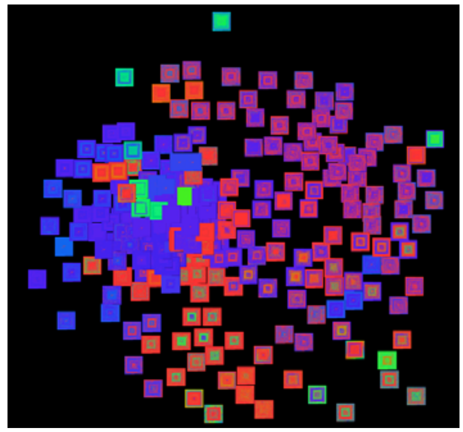
	d1	d2	d3	d4
d1	0	0.1	0.2	0.7
d2	0.1	0	0.3	0.6
d3	0.2	0.3	0	0.7
d4	0.7	0.6	0.7	0

Multi-Dimensional Scaling

Value and Relation Display

Features:

- Explicitly conveys data values without dimension reduction
- Explicitly conveys dimension relationships
- Provides a rich set of interaction tools

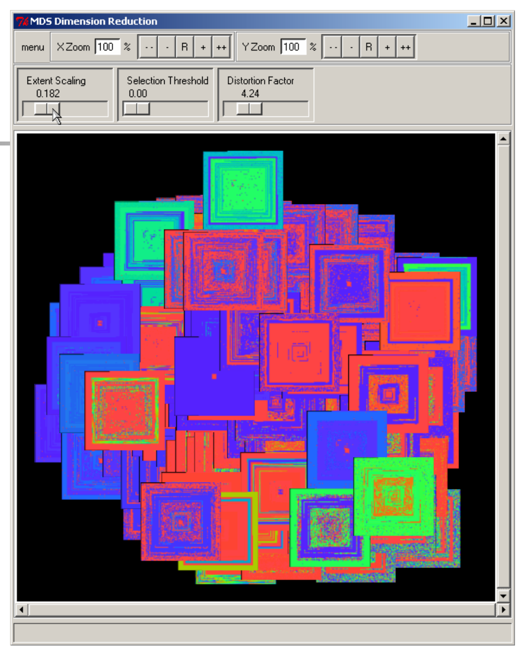


SkyServer dataset: 361 dimensions, 50,000 data items

47

Overlap Detection and Reduction

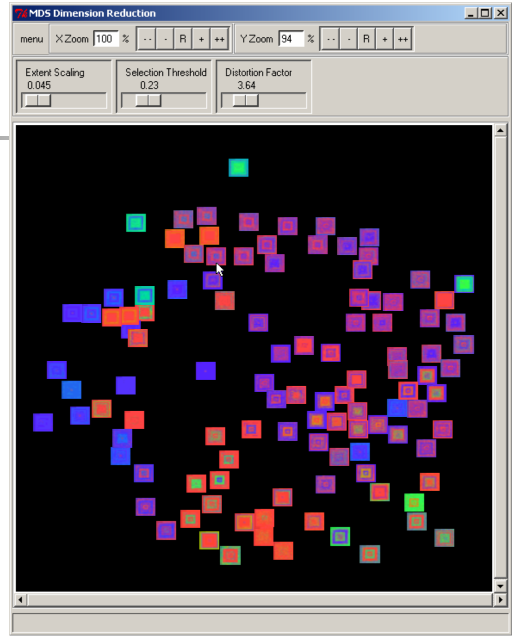
- Extent Scaling
- Dynamic Masking
- Zooming and Panning
- Showing Names
- Layer Reordering
- Manual Relocation
- Automatic Shifting



SkyServer Dataset

48

Distortion



Goal:

- Focus-within-context

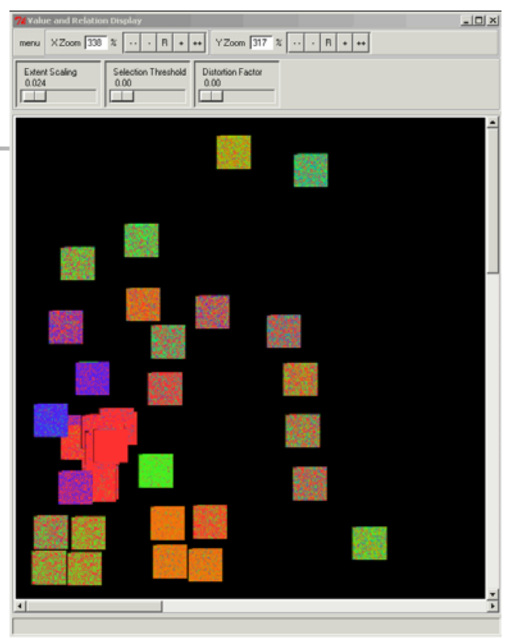
Features:

- Enlarges clicked glyphs
- Keeps size of other glyphs

SkyServer Dataset

49

Data Item Reordering



Pixel-oriented techniques:

- Data item ordering is critical

VaR display:

- Initial display
- Manual reordering

Census-Income-Part Dataset:
42 dimensions, 20,000 data items

50

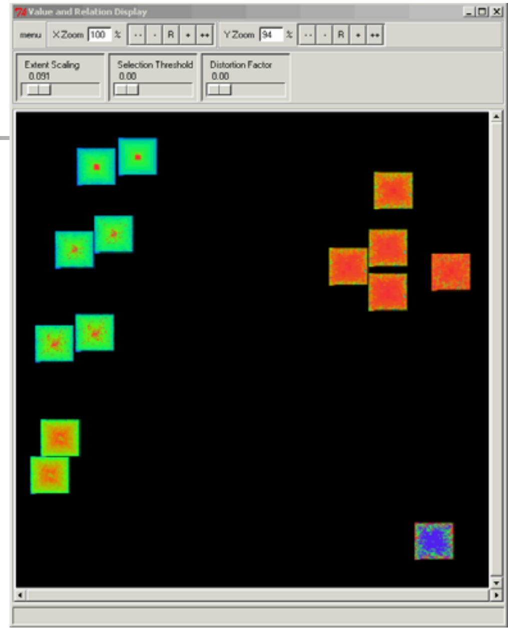
Comparing

Goal:

- Compare base dimension with all others

Feature:

- Coloring by value difference of dimensions being compared



AAUP Dataset:
14 dimensions, 1,131 data items

51

Selection

Goal:

- Select dimensions for further interaction or visualization

Selection tools in VaR display:

- Manual selection - flexibility
- Automatic selection - efficiency
  - Select related dimensions
  - Select unrelated dimensions

52

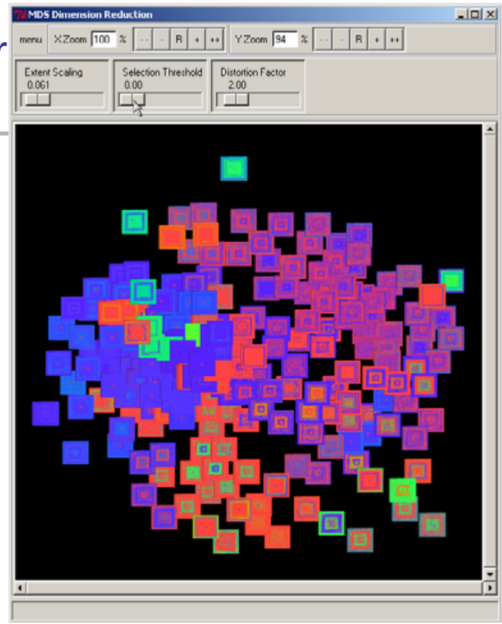
Automatic Selection for Unrelated Dimensions

Input:
- A base dimension
- “Related” threshold

Output:
- Dimensions covering major data variance

Algorithm:
- Iteratively select unrelated dimensions and filter related dimensions

Related work:
- Maximum subspace [*MacEachren*:03]



SkyServer Dataset

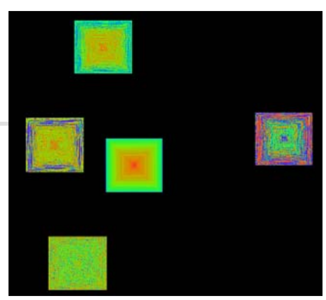
53

Scale to Large Datasets

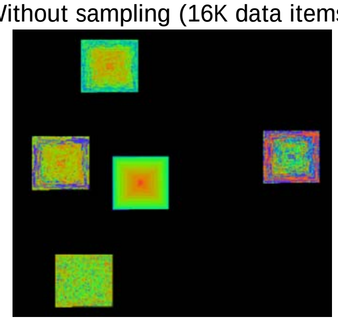
Store glyphs as texture objects
- Extent scaling and relocating: resize, relocate texture objects ☺
- Reordering and recoloring: regenerate texture objects

Use random sampling
- Users interactively set threshold
- Random sampling is triggered automatically

Out5D Dataset



Without sampling (16K data items)



With sampling (5K data items)

54

Discussion

Is 2D MDS the only approach to layout dimensions?

- 3D MDS, SOM, ...

Is pixel-oriented technique the only choice for generating dimension glyphs?

- Histogram, Scatterplot, ...

Is correlation the most informative relationship between dimensions?

- User defined distances

Possible Applications of VaR Display

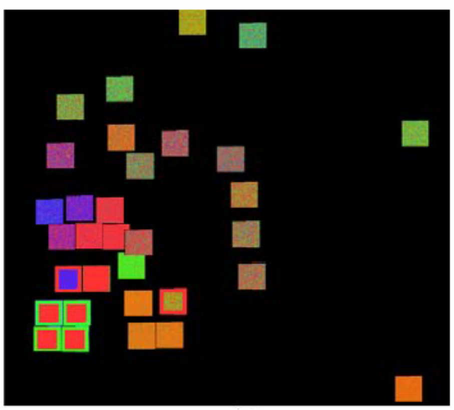
- Interactively exploring high dimensional data
  - Revealing data item relationships
  - Revealing dimension relationships
- Guiding automatic data analysis
  - Assessing results
  - Manually tuning parameters
- Human-driven dimension reduction
  - Constructing subspaces using selection tools
  - Visualizing subspaces in VaR or other displays

Case studies

- AAUP Dataset: 14 dimensions, 1,131 data items
- Census-Income-Part Dataset: 42 dimensions, 20,000 data items
- OHSUMED dataset: 215 dimensions, 298 data items
- SkyServer dataset: 361 dimensions, 50,000 data items

57

Example



Census-Income-Part dataset: (bottom left) a group of dimensions recording people's migration and moving status in the last year.

58

VHDR vs. VaR

Similarity:

- Both scale to high-dimensional datasets
- Both visually reveal dimension relationships

Differences:

- VaR visualizes dimension relationships in the same display as data values
- VHDR visualize dimension relationships and data values in separated displays

59