

Visual Data Mining

Chidroop Madhavarapu
CSE 591: Visual Analytics

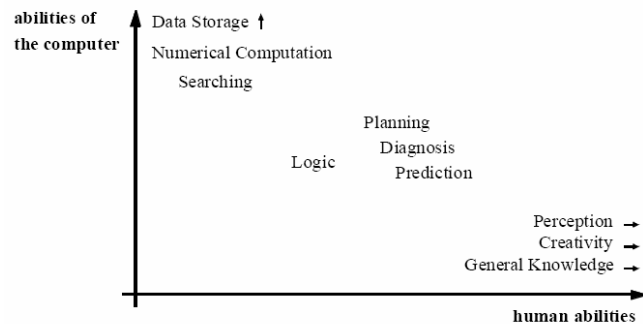
Motivation

Visualization for Data Mining

- Huge amounts of information
- Limited display capacity of output devices

Visual Data Mining (VDM) is a new approach for exploring very large data sets, combining traditional mining methods and information visualization techniques.

Why Visual Data Mining



Integration of visualization and data mining :

Visual Data Mining approaches fall under 3 categories:

- Data Mining process visualization.
- Data Mining result visualization.
- Interactive Visual Data Mining.

Data Mining process visualization

Visualization techniques are used to support Data Mining.

Ex: When required to handle large amount of multidimensional data in the format of Data Tables or relational databases. (Parallel Coordinates, scatter plots etc.)

Data Mining result visualization

To visually convey the results of Mining tasks, such as clustering or classification, to enhance user interpretation.

Examples include Scatter plots, Box plots, BLOB and H-BLOB clustering algorithms, Decision trees, Association Rules,

Interactive Visual Data Mining

- Rather than using Visual Data exploration and analytical mining algorithms as separate tools, a stronger DM strategy would be to tightly couple the visualizations and analytical processes into one DM tool.
- Using visualization tools in the data mining process to [help users make smart data mining decisions](#).
- Examples include the Control project, OptiGrid, PBC (Perception Based Classification).

V-Miner: Using Enhanced Parallel Coordinates to Mine Product Design and Test Data

INTRODUCTION

- V-Miner : Multivariable visualization tool.
- Designed for Mining product design and test data.
- New technique based on Parallel coordinate visualization.
- Goal is to discover useful knowledge from mobile phone testing data that can be used to provide feedback to the design engineers.

Design Process for consumer electronics.

- Engineers design specific sections of phone based on previous successful designs, new product specs, design simulations etc.
- Prototypes are built
- Functional tests are performed on prototypes. If the requirements are not met, start with next design cycle (from step 1).

Above steps are repeated until design meets the specification. Then the phone is released to the NPI team for volume manufacturing.

- For a new product, number of iterations of design revisions should be coordinated.
- 100's of variables involved which are changed/tested in the different revisions.
- V-Miner is used to reduce engineering costs, design defects by mining useful knowledge from the test data .

THE DATA

After each design change, all test variables are measured. Each variable takes numerical values and has the following properties:

- Has an *upper limit* and *lower limit*. If a value does not fall in this range, its unacceptable.
- Has an ideal value called the *target value*.

SAMPLE TEST DATA

Each change is a new design. Data is a sequential set.
Subsequent changes are based on earlier changes.

With the testing data, designers are interested in :

- Significant changes in variables with design change.
- Cause of these changes.
- Stable variables whose values are not affected by design changes.

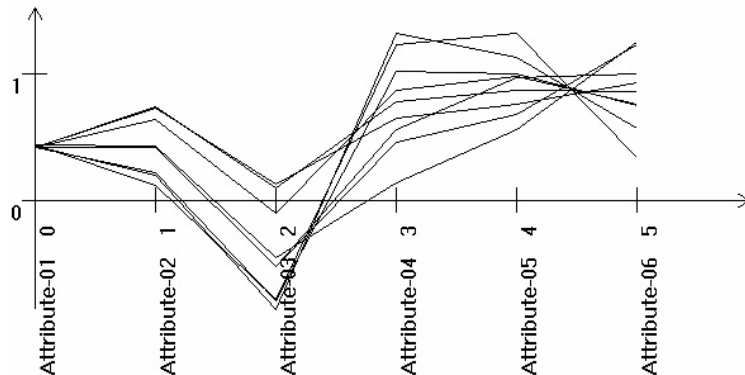
Test variable	Target value	Lower limit	Upper limit	Value unit	Value for change 1
Variable 1	0	- 4	4	Voltage	0.9
Variable 2	50000	50000	55000	HZ	495830
.....

Using Traditional Mining algorithms is not adequate here because,

- Due to large number of variables, association rule mining generates too many rules.
- Decision trees does not find all interesting patterns, but only subset of the patterns.

To solve the problem, we can use parallel coordinates which give an intuitive view to the underlying data.

Parallel Coordinates Overview



Problem with the traditional parallel coordinates technique

- Does not consider the sequence in which the data was generated. Sol: Add a sequence component to the traditional parallel coordinate visualization. -- Add *trend figures*.
- Does not consider the ordering of the variables. Sol: A querying and sorting tool is implemented to enable users to issue queries and rearrange the axes accordingly.

So, design an Enhanced Parallel Coordinates system.

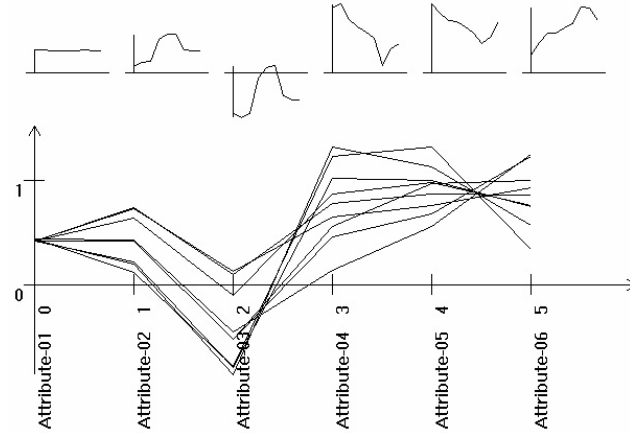
TREND FIGURES

Order of data records is of high significance, as it might reveal sequence dependent relations.

Extend the existing system by adding a additional graph for each variable above its coordinate.

Thus it is possible to quickly see variables that change in similar ways by comparing the trend figures.

TREND FIGURES



QUERYING AND SORTING

Allows user to query shapes based on approximate pattern matching.

Two main types of pattern: Value change pattern & Failure pattern.

Value change pattern indicates how a variable's value changes over different design changes. **up :3 down: 1 stable: 2**

Example: 3312

Failure pattern indicates if the value falls within the upper and lower limit after the design modification.

F: failure O: ok Example: OOOFF

String comparison is more convenient and intuitive for human users.

Ordering of the variables in parallel coordinate visualization is done according to the comparison results.

Need for Data Mining

Goal for the application is to enable engineers at Motorola to identify the following:

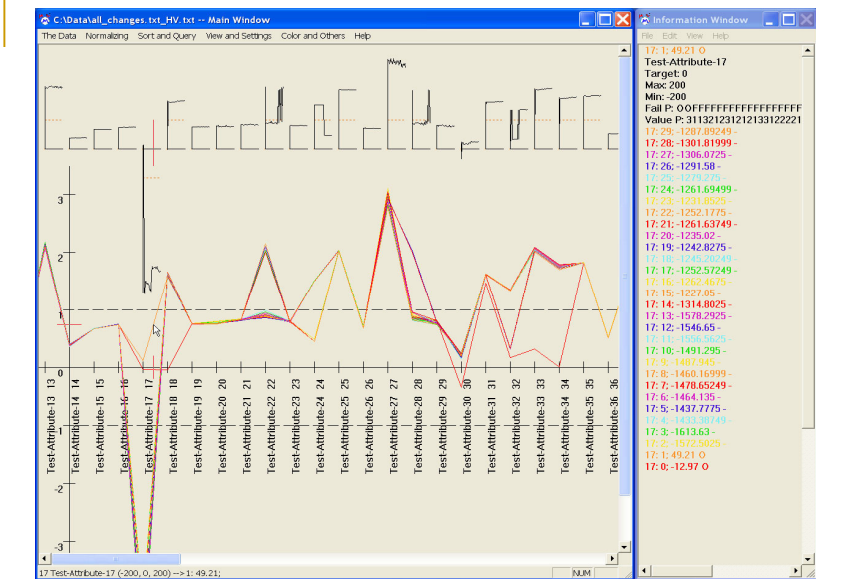
- Variables that show prominent changes in their values after some design changes.
- Stable variables that aren't affected by the design changes.
- Failure patterns of variables that failed after certain design changes.
- Variables that have similar value change patterns.

DATA NORMALIZATION

Variables whose values are out of range are normalized to either larger than 1 or less than -1. Normalized values close to 0 are the ones close to the target values.

Procedure normalization (value, min, max, target)

```
// return value stores in: normalized_value
if ((value >= min) && (value <= max)) then
    normalized_value = (value - target) / (max - min);
else
    if (value > max) then
        normalized_value = (value - target) / (max - min) + 1;
    else // value < min
        normalized_value = (value - target) / (max - min) - 1;
    end-if
end-if
```



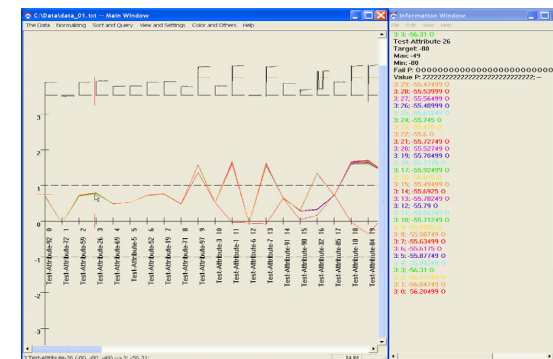
KEY FEATURES

- Data in different designs are visualized using different colors.
- For each variable, a trend figure is drawn on the top of the screen
- User can identify significant characteristics from visualization .
- User can easily identify which variables are out of range or within the range (ex 19, 20).
- Variables that behave similarly from the trend figures (ex 33, 34).
- Some variables have stable values over all design changes (ex 15).

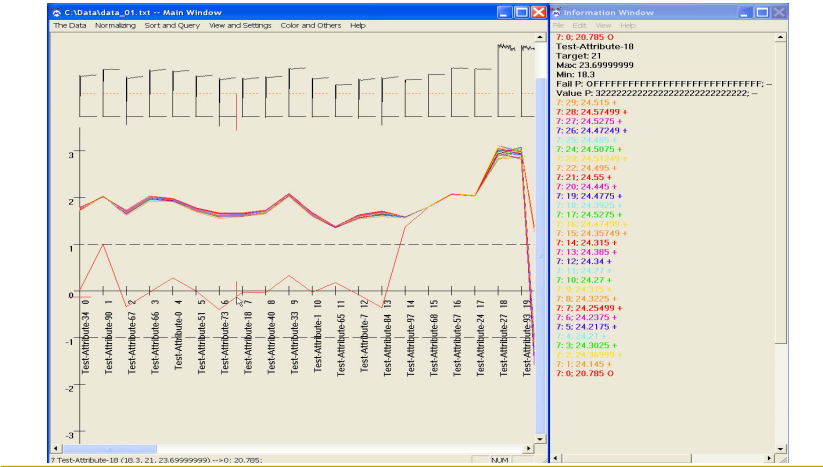
In classical parallel coordinate visualization, overlapping lines significantly hinders visualization. Trend figures mitigate the problem.

STABLE VARIABLES

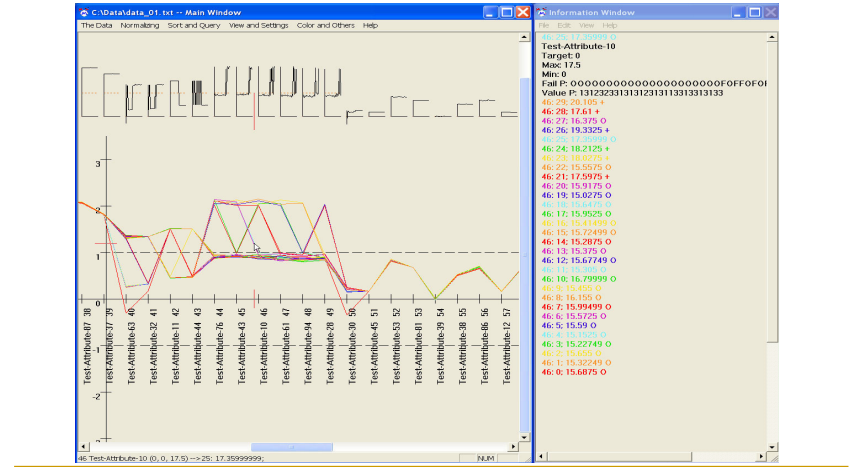
If the user is interested in identifying stable variables, he can issue a '222...' query on the value change pattern. Variables are ordered with the stable variables appearing first.



Test variables that failed after the first design change.



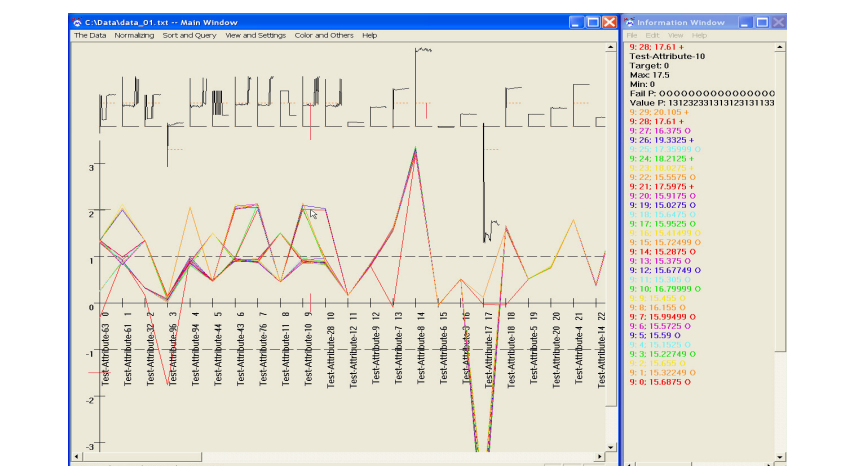
Variables affected by many components



Consistent Failures

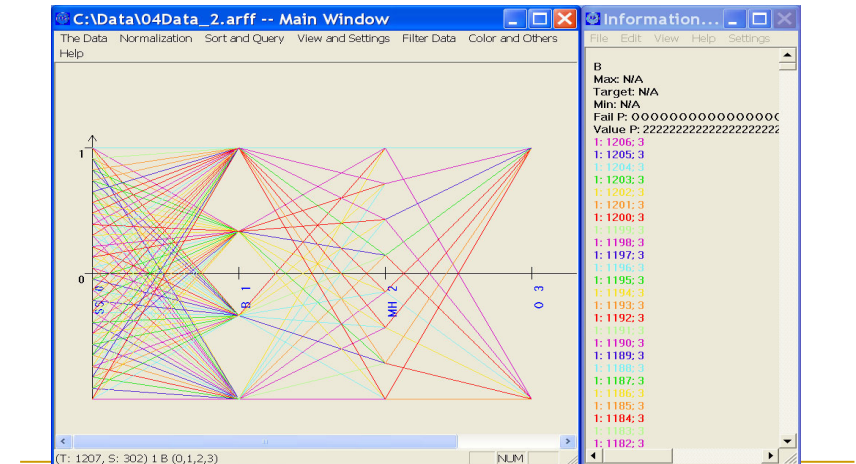


Result of a 'FO' Pattern



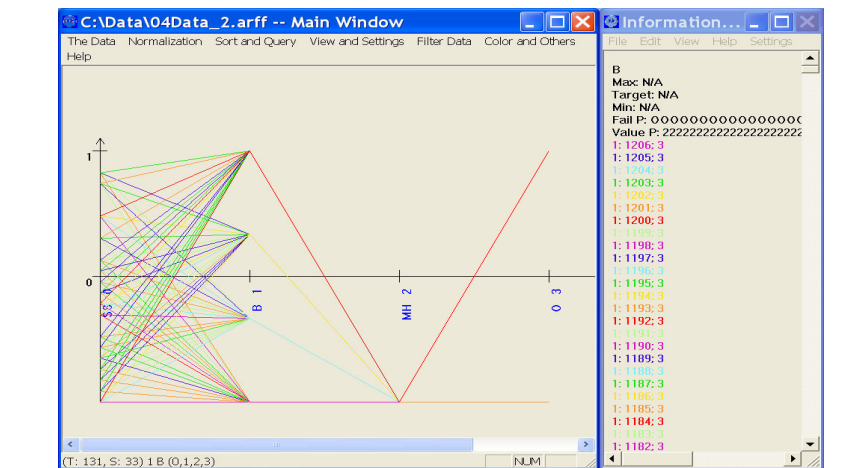
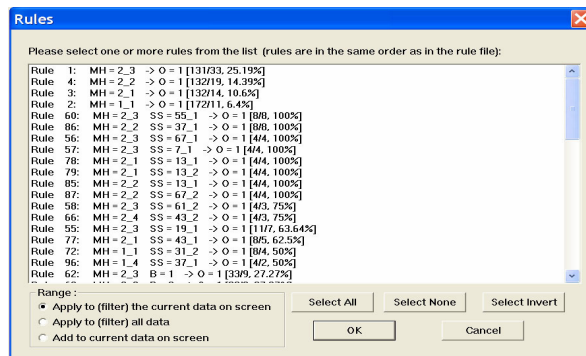
Initial Visualization of a data set

- V-Miner can be used with the existing data mining tools. Previously mined rules can be used to filter the data in the visualization.
- V-miner does not act simply as a tool that filters data, instead provided opportunity to the user to interact with the data visually.



Visualization of data after filtering using rules

Engineers have a set of rules from a data mining tool, they are loaded into V-Miner and user can select rules from here.



CONCLUSIONS

- This visualization system significantly speeds up the data mining process.
 - V-Miner is able to find knowledge that cannot be found by other tools like correlation between variables, failure patterns in sets of multiple variables.
 - Engineers can use V-Miner and their favorite mining tools together and recursively to mine for finer details.
-

Interactive Data Analysis : The CONTROL Project

Joseph M Hellerstien, Ron Avnur, Andy Chou, Chris Olston, Vijayshankar Raman.
University of California, Berkley

Data Analysis

Objective: Obtain unknown information.

- Is an Iterative process.
 - Complex process involving multiple, time consuming steps.
-

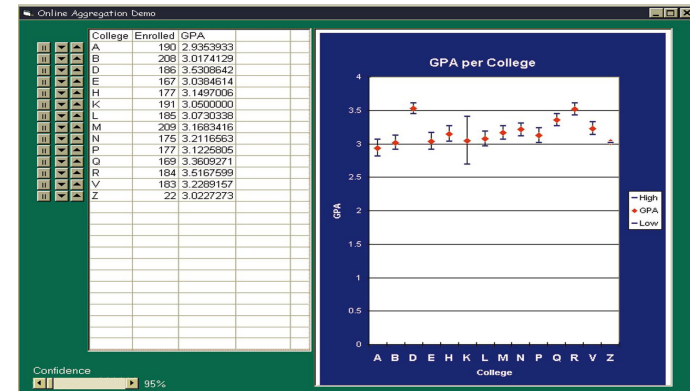
-
- Batch (Current systems) Vs Online Processing (CONTROL - **C**ontinuous **O**utput and **N**avigation **T**echnology with **R**efinement **O**nline).
 - Black Box Vs Crystal Ball.
 - Quality & Accuracy Vs Interactive response times.
-

Online Aggregation

- Relational Databases Vs Online query processing.

```
SELECT college, AVG(grade)
FROM enroll
GROUP BY college;
```

- Relational DB: Partition, Calculate, Return.
- Online aggregation system:



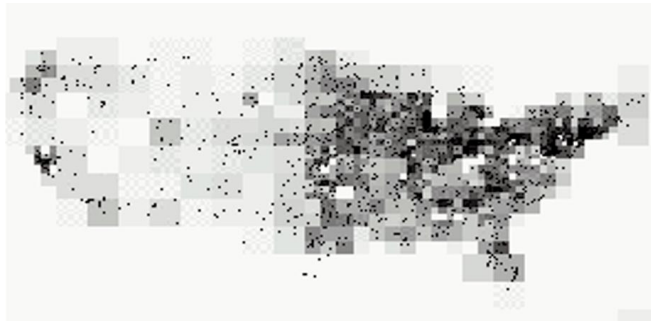
Online Data Visualization: CLOUDS

GOAL: Make Visualization more interactive by quickly displaying an accurate approximation of the final image.

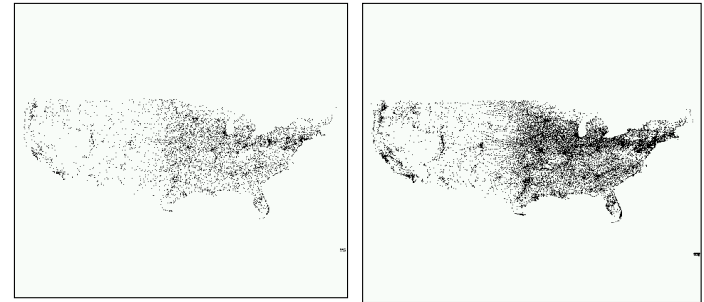
Partially completed visualization of US cities without CLOUDS:



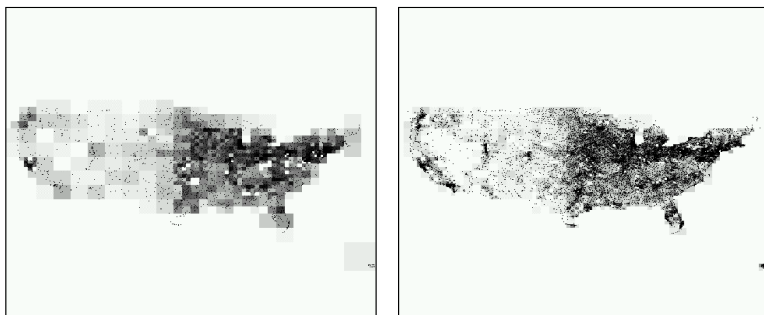
Partially completed visualization of US cities with CLOUDS:



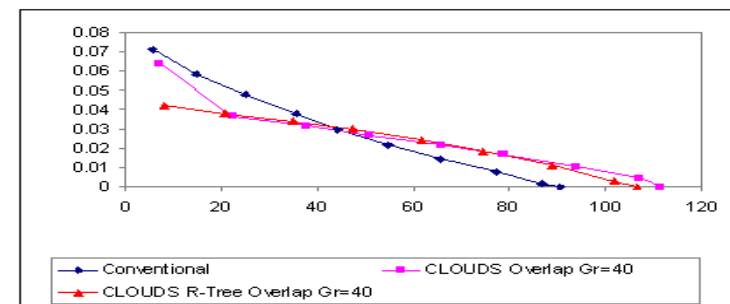
US cities with conventional algorithm after 25 and 65 seconds:



US cities with Clouds algorithm after 25 and 65 seconds:



Graph of Mean Squared error over time. Clouds have lower error over non clouds.



Sampling from Multiple Joins : Ripple Joins

- Classical join algorithms scan large portion of input before they return records.
 - Ripple join algorithm can start returning output immediately upon invocation.
-

Ripple Join: Operation

Assume ripple join of relations R and S

- Select random tuple r from R.
 - Join with previously selected S tuples.
 - Do random select s from S.
 - Join with previous R tuples.
 - Join r and s.
-

Ripple Join

In each matrix in the figure, the R axis represents tuples of R, the S axis represents tuples of S, each position (r, s) in each matrix represents a corresponding tuple in R x S.

The “square” version of this ripple join samples from R and S at the same rate.

Ripple Join: Square Two-Table Join

R
S X

Ripple Join: Square Two-Table Join

R
S X X
X X

Ripple Join: Square Two-Table Join

R
S X X X
X X X
X X X

Ripple Join: Square Two-Table Join

R
S X X X X
X X X X
X X X X
X X X X

CONTROL Today

- Control Algorithms is used in several freeware and commercial systems.
 - Online aggregation techniques is integrated into the DB2 Universal Database.
 - CLOUDS is implemented in Berkley's Tioga Datasplash visualization system.
-
