Visualization of Multi Parameter Hierarchical Data Using Automatic Dominant Parameter Determination Technique

Takeru Kiyoshi[*1] **Takayuki Itoh**[*1*3*4] **Koji Koyamada**[*2]
Koji Sakai[*2] **Takeshi Iwashita**[*1] **Masanori Kanazawa**[*1]
*1 Academic Center for Computing and Media Studies, Kyoto University
*2 Center for the Promotion of Excellence in Higher Education, Kyoto University
E-mail: kiyogou@nifty.com, itot@computer.org

Abstract

This paper presents a visualization technique of hierarchical data which each leaf and non-leaf node has multi parameters. The technique determines the dominant two parameters from the multi parameters, by applying response surface technique. By assigning the two parameters to horizontal and vertical axes of display spaces, the technique represents the dependency among the dominant parameters of hierarchical data. The technique applies HeiankyoView, a visualization technique for large-scale hierarchical data. The paper introduces some visualization results proofing the effectiveness of the presented technique, and a scientific application that the presented technique is to be effectively used.

1. Introduction

Multi parameter hierarchical data is a common data structure. Figure 1 represents the definition of multi parameter hierarchical data in this paper. The paper supposes that leaf and non-leaf nodes have each own multi parameters. Such data structure widely appears in our real life. For example, company employee data usually form hierarchical organization, and each employee contains multiple values, such as duration of work, salary, and scores of English examinations. Network computer data usually form hierarchical network structures, and each computer contains multiple measurement values, such as numbers of packets in a constant time, number of partner computers of communications, and numbers of communication errors.

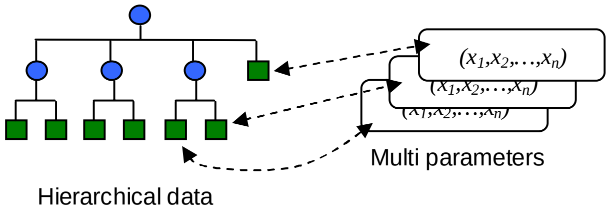


Figure 1. Definition of multi parameter hierarchical data.

Generally scientific simulations need adjustments of multi parameters to obtain realistic solutions, and therefore information visualization techniques are useful for understanding dependency between the parameters and results of simulations [1].

Some of authors have experiments of multi parameter data of scientific simulations. Here, let the error of the simulation y as $y = S(x_1, x_2, ..., x_n) - C$, where S is the solution of the simulation, x_1 to x_n are input parameters of the simulation, and C is the result of real, ideal experiments or observations.

It is very important to discover the values of the parameters that lead the best solutions of the simulations. Some of authors have presented a technique to discover such values of parameters in reasonable computation time [2]. Given the initial ranges of the parameters, the technique invokes the simulation with several sets of sample values of input parameters, and finds preferable smaller ranges of the parameters. It then invokes the simulation with the new sample values in the smaller ranges, and again finds the preferable, smaller ranges of the values. Repeating this procedure, the technique arrives at enough narrow ranges of values and finds the best values that lead the smallest error of the simulation. During the procedure invoking the simulation m times, the technique obtains a set of simulation results $\{y_1,..., y_m\}$ forming hierarchical data according to ranges of the parameters, and each simulation result y_j has values of the parameters, x_1 to x_n.

Some of authors have also presented HeiankyoView [2], which represents large-scale hierarchical data as a collection of icons and nested rectangles. Such hierarchical data visualization technique is especially useful to intuitively understand the distribution of simulation results. Our target in this paper is the understanding of dependency of simulation results y with the input parameters x_1 to $x_n,$ by using HeainkyoView. Since every parameter do not always strongly affect to the simulation results, determination of dominant parameters is very important to realize the effective visualization.

This paper presents a technique for visualization of multiple parameter hierarchical data. The technique first determines the most dominant two parameters, x_i and x_j, by using response surface method. The method generates surfaces interpolating the parameters x_1 to x_n and the simulation results y as a trial, and determines the two parameters, x_i and x_j , which are most dominant to y. The technique then visualizes the hierarchical data by HeiankyoView, while it maps the values of x_i and x_j as the values of horizontal and vertical coordinates of the display. This paper represents the experimental results of the presented technique, and introduces a scientific application that the presented technique is to be applied.

2. HeiankyoView

HeiankyoView [3] is a new hierarchical data visualization technique, which represents the data as a set of icons and nested rectangles. This representation style is very similar to Data Jewelry Box [4] presented by Itoh et al., but HeiankyoView

*3 IBM Research, Tokyo Research Laboratory

*4 Department of Information Sciences, Ochanomizu University applies completely different algorithm to place the icons and rectangles onto display spaces. Figure 2 shows an example of the visualization of hierarchical data by HeiankyoView. The name of the technique comes from Heiankyo, an ancient palace in Japan. It is well-known that land arrangement of Heiankyo was very well-organized by grid-like blocking.

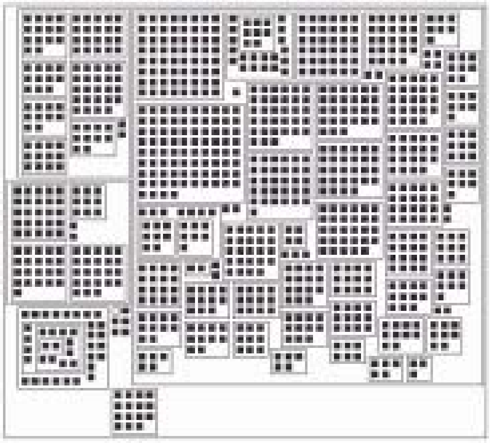


Figure 2. An example of visualization of hierarchical data by HeiankyoView.

Such hierarchical data visualization technique is very useful and therefore it has a lot of applications. Actually Itoh et al. have presented various applications of the hierarchical data visualization technique, including access analysis of Web sites [5], monitoring of distributed computing environments [6], and monitoring of network intrusions [7].

HeiankyoView places the icons and rectangles onto display spaces one-by-one, according to the following two conditions:

[Condition 1] Icons and rectangles under a same parent never overlap each other.

[Condition 2] Icons and rectangles are placed where it minimizes the display area after the placements.

The above conditions do not directly control the positions of the icons and rectangles on the display, but it is much more useful if an additional condition controls their positions. As shown in Figure 3, this paper applies the following additional condition:

[Condition 3] Given ideal positions of the icons and rectangles on the display, they are placed where are close to the ideal positions.

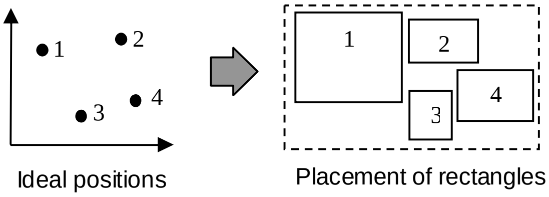


Figure 3. Ideal positions of icons and rectangles.

Here let us consider of calculating the ideal positions using the selected two parameters, x_i and x_j. Our target is calculating ideal horizontal coordinates of icons and rectangles from values of the parameter x_i as ideal, and ideal vertical coordinates of them from values of the parameter x_j.

Let us also suppose that the values of y are represented as heights of nodes by HeiankyoView. Since the selected two parameters x_i and x_j strongly affect to y, smoother variation of heights of nodes are shown, rather than when the two parameters are randomly selected, as shown in Figure 4. Proper selection of the two parameters will therefore strongly help to understand the dependency of values of y with the dominant parameters. Next section describes a technique to automatically select the dominant two parameters x_i and x_j.

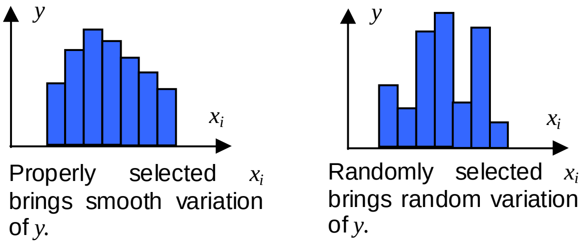


Figure 4. Variation of values. Proper section of parameters leads to understand the dependency between x_i and y.

3. Dominant Parameter Determination

Suppose that each parameter x_i has k values, and the result y also has k values. Here the technique generates a surface interpolating the above values, applying least square method. The surface is called as "response surface." This is a general technique introduced in many textbooks in numerical methods.

Response surface is an interpolation of equation of n variables and response y. Assuming that the surface is a quadric polynomial surface, it is represented as the equation (1).

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1, j\le i}^{n} \beta_{ij} x_i x_j \quad ..(1)$$

When k sample points and responses $y_i\,(i = 1,\ldots,k\,)$ are given, p variables are represented as equation (2).

$$Y = X\beta + \varepsilon \quad ..(2)$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{k1} & x_{k2} & \cdots & x_{kp} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix}.$$

Unbiased estimator b of the coefficient β is represented as equation (3):

$$b = (X^T X)^{-1} X^T y \quad ..(3)$$

The equation (3) is used to calculate the coefficients β. After calculating them, the response surface technique tests validity of β.

Generally coefficient of determination R^2_{ad}, calculated according to the equation (4), is used for the evaluation of β:

$$R^2_{ad} = 1 - \frac{SS_R \big/ (k-p-1)}{S_{yy} \big/ (k-1)} \quad ..(4)$$

where

$$SS_R = {\beta_b}^T X^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \text{, and}$$

$$S_{yy} = y^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

If the value of R^2_{ad} is too small, then the technique determines that the given *m* sets of values are too random or noisy to obtain an adequate surface. In this case our implementation gives up selecting the dominant parameters.

Otherwise, the technique tests each value of β to determine the dominant two parameters. It applies t-test to calculate the reliability of the coefficients, and eliminates some coefficients which bring small *t* values. By repeating the calculation and the elimination, the technique reduces the number of parameters and improves the reliability of the equation (1). Here *t* value of j-th coefficient is calculated by the equation (5), where σ is the maximum likelihood estimate of the distribution of β, and C_{jj} is jj-element of the matrix $(X^TX)^{-1}$.

$$t = \frac{\beta_j}{\sqrt{\hat{\sigma} C_{jj}}} \quad ..(5)$$

In equation (1), parameters multiplying coefficients, which have larger *t* values, are dominant to the response *y*. We suppose that such parameters are preferable to use for calculating ideal horizontal and vertical coordinates of display spaces.

4. Experiments

In this experiment we used a small hierarchical data whose elements contain three parameters, x_1 to x_3, and another value *y* dependent from the three parameters. Figure 6 introduces the visualization results, where:

- Color of nodes denotes the values of parameters x_i used for calculating horizontal coordinates. The colors just proofs if x_i adequately brings the ideal x-coordinate values of nodes. Figure 6 (Center) and (Right) denotes that the hue of nodes smoothly varies from blue to red along the horizontal axis of the display space, and therefore it proofs that x_i adequately brings preferable x-coordinate values. We confirmed that x_j also adequately brings preferable y-coordinate values.
- Height of nodes denotes the values of *y*. We can determine that better parameters are selected if the height of nodes smoothly varies in the image.

Figure 6 (Left) denotes the result without calculating ideal positions. Figure 6 (Center) denotes the result with ideal positions calculating by the parameters selected by the presented technique. Figure 6 (Right) denotes the result with ideal positions calculated by the randomly selected parameters. The result using the presented technique shows smoother variation of heights of nodes, and therefore it proofs that the technique succeeds to select the proper parameters x_i and x_j dominant to the value *y*.

Figures 7 and 8 denote similar results, proofing that the technique succeeds to select the proper parameters.

5. Applications

Currently some of authors are applying this technique to the visualization of results of cell simulation [8]. The simulation calculates the time sequence of active voltage of cells, inputting enormous number of parameters including initial potential of ions, such as K+, Ca+, and Na+. The simulation results are often compared with measured real values. Figure 5 is the window capture of the cell simulation system. Left part of the window lists enormous number of parameters, and right part draws multiple time sequence values.

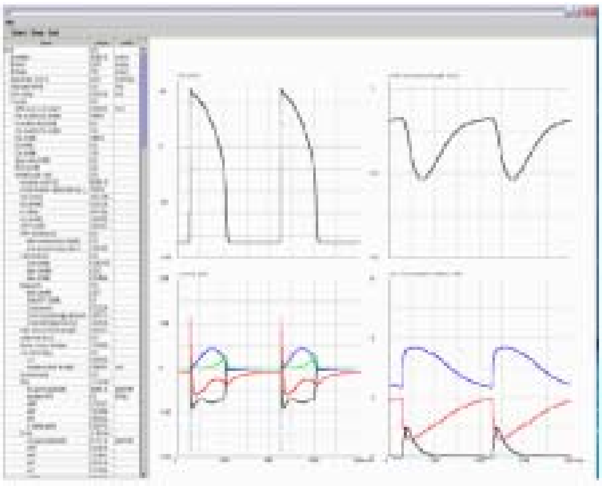


Figure 5. Window capture of a cell simulation system.

Discovery of dependency among the simulation results and the input parameters are a difficult problem and strong interest of scientists. The presented visualization technique is extremely useful for this purpose. Let us represent the error of simulation result against the measured real value as *y*, and the parameters as x_i. To understand the dependency the simulation should be repeated with changing the parameters, to obtain the collection of error values. The cell simulation system preserves the collection of error values and parameters as hierarchical data, by categorizing the error values according to ranges of parameters, and treating the error values as leaf-nodes of a tree structure, as described in Section 1 [2]. The presented technique can be used to determine dominant parameters to realize smooth visualization of the collection of error values. We are going to apply the presented technique for the effective visualization of results of the cell simulation.

Also, we would like to apply the technique to various multi parameter hierarchical data in addition to cell simulation in the near future.

6. Conclusion

This paper presented a technique of determination of dominant parameters for effective visualization of multi parameter hierarchical data. The technique applies response surface technique that leads the surface interpolating collections of parameters and responses. The response surface technique first determines the validity of the equation representing the surface, and then reduces the number of parameters. Finally it tests the parameters to determine the most dominant two parameters.

The paper introduced some visualization results HeiankyoView, which assigns the values of the dominant parameters to two axes of display spaces. The results proofed that the presented technique realized smooth representation of the multi parameter hierarchical data.

Future works of the technique include:

- numerical evaluation of the visualization results,
- combination with other visualization techniques in addition to HeiankyoView, and
- application to various multi parameter hierarchical data in addition to cell simulation.

References

[1] Obayashi S., Sasaki D., Visualization and Data Mining of Pareto Solutions Using Self-Organizing Map, Second International Conference of Evolutionary Multi-Criterion Optimization, pp. 796-809, 2003.

[2] Koyamada K., Sakai K., Itoh T., Parameter Optimization Technique Using the Response Surface Methodology, IEEE Engineering in Medicine and Biology Society, 2004.

[3] Itoh T., Koyamada K., HeiankyoView: Orthogonal Representation of Large-scale Hierarchical Data, International Symposium on Towards Peta-Bit Ultra Networks (PBit 2003), pp. 125-130, 2003.

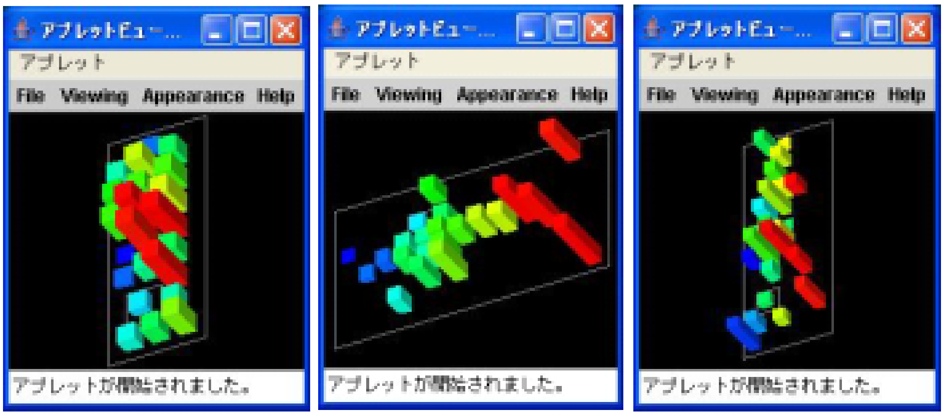
[4] Itoh T., Yamaguchi Y., Ikehata Y., Kajinaga Y., Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, IEEE Transactions on Visualization and Computer Graphics, Vol. 10, No. 3, pp. 302-313, 2004.

[5] Yamaguchi Y., Itoh T., Ikehata Y., Kajinaga Y., Interactive Poster: Web Site Visualization Using a Hierarchical Rectangle Packing Technique, IEEE Information Visualization Symposium, 2002.

[6] Yamaguchi Y., Itoh T., Visualization of Distributed Processes Using "Data Jewelry Box" Algorithm, CG International 2003, pp. 162-169, 2003.

[7] Itoh T., Takakura H., Sawada A., Koyamada K., Visualization of Network Intrusion Detection Data Applying a Hierarchical Data Visualization Technique “Heiankyo-View'”, Eighth International Symposium on Recent Advances in Intrusion Detection, submitted.

[8] Sarai N., Noma A., simBio: A Java Package for Biological Simulation, 5th International Conference on System Biology, p373, 2004.

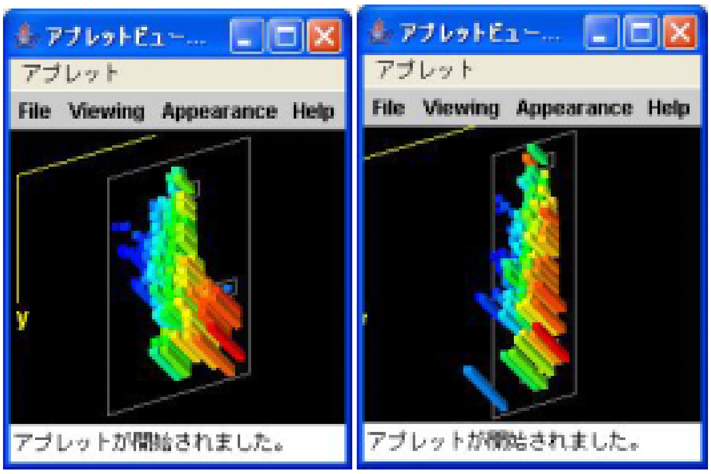


アプレットが開始されました。

アプレットが開始されました。

アプレットが開始されました。

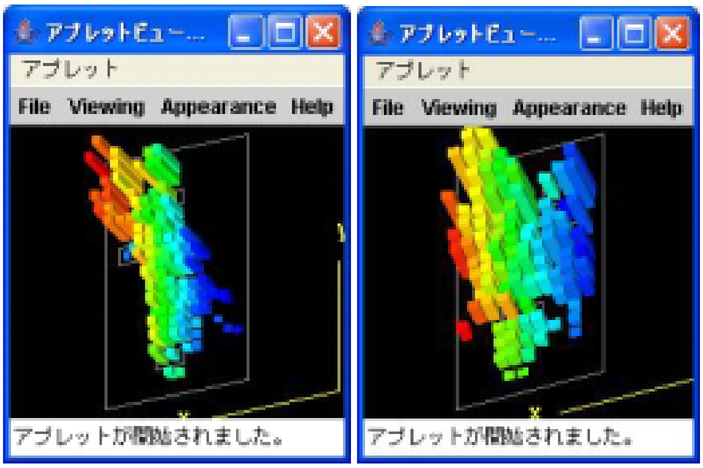
Figure 6. Experimental results (1). (Left) Ideal positions are not calculated. (Center) Parameters are selected by the presented technique. (Right) Parameters are randomly selected.



アプレットが開始されました。

アプレットが開始されました。

Figure 7. Experimental results (2). (Left) Parameters are selected by the presented technique. (Right) Parameters are randomly selected.



アプレットが開始されました。

アプレットが開始されました。

Figure 8. Experimental results (3). (Left) Parameters are selected by the presented technique. (Right) Parameters are randomly selected.