The Study of Past Working History Visualization for Supporting Trial and Error Approach in Data Mining

Kunihiro Nishimura[1] and Michitaka Hirose[2]

[1] Research Center for Advanced Science and Technology, The University of Tokyo
4-6-1, Komaga, Meguro-ku, Tokyo, 153-8904, Japan
[2] Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{kuni,hirose}@cyber.rcast.u-tokyo.ac.jp

Abstract. Scientists in the data mining field are constantly faced with the challenge of finding useful information from a huge amount of information. We have to analyze the data until we can get the appropriate information. We have to select one part of the data, compare them against each other, or arrange them in certain order. This approach is also known the trial and error approach. A trial and error approach requires the users' judgment, for example, to correctly set certain parameters; it is an approach that place importance not only to the end result, but also to the process in achieving the end result. In this paper, we propose visualization methods to visualize past working history for supporting trial and error approach in data mining. We use our methods to visualize web browsing logs and data browsing logs in genome science fields.

Keywords: Information Visualization, Past Working History, Web Visualization.

1 Introduction

We usually have to shift through a vast amount of information that is available on the internet or from knowledge databases when we want to find a specific answer to our question. If we know the target keyword, we can perform the search and find the solution almost immediately. However, if the subject is more ambiguous, we might not have a target keyword immediately. We have to perform the search by specifying several target keywords, check the result and refine the query string iteratively. In short, we have to do the trial and error approach.

When conducting a research by trial and error, we tend to forget previous trials. However, there is a possibility that the previous trials' results are better than the current result. Also, there are times when we do the same trial because we forgot that it has already been conducted in the past. In this study, we developed visualization techniques to support trial and error approach for data mining.

Our proposed visualization method visualizes the working process as well as the results. Because the users are able to glance at the parameters of trial and errors in the past, they are able to deduct the relationship between the trial result and the parameters used in that trial.

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2007, LNCS 4557, pp. 327–334, 2007.

2 Information Visualization for Supporting Data Mining

Information visualization is the use of computer-supported, interactive, visual representations of abstract data to amplify cognition [1]. Visualization with spatial structures is called scientific visualization. On the other hand, information visualization is a term applied to visualizing information without spatial structures. For example, web networks, documents, and directory structures are targets of information visualization [2]. With adequate information visualization, it is easy to carry the information to the users. Information visualization is an effective method to grasp a complete view of the data.

In data mining, we want to grasp, understand, and interpret the target data. We manipulate, compare, calculate the data, and get the information we want. In order to optimize the search result, data mining process often requires user interaction. When the data mining result are visualized, it is easier for users to understand the result. Thus, we believe that the visualization of data with interaction support can enhance users' data mining process.

Users do trial and error approach in data mining when they want to get specific information in the data. Visualization of the manipulated result is called "visualization of trial". Visualization methods of trial often include displaying various kinds of graphs. Thus, there are a lot of visualization methods of trial. During data mining process, users manipulate of the data (this is called a trial), and check the result by visualization of the trial. Then users interpret them and decide next manipulation. Users repeat this trial (manipulation, check the visualization result, and interpretation and decision) until they can get suitable information.

In addition, working process is as important as trials in data analysis. A final result is often based on the previous trials. Relationship between current trial results and previous results are important because trial and error approach often depend on accumulation of the user's previous trials' result. That is, it is important to display the "result" of trials but also "process" which means "past working history" and context of analysis in data mining.

In this paper, we focus on visualization of past working history for supporting trial and error approach in data mining.

3 Visualization of Past Working History

As we have stated in previous section, it is important to visualize working result but also to visualize working history. In this section, we discuss visualization of past working history.

Past working history consists of time and work process. When data is visualized and a user manipulates it, data has spatiality. Thus, past working history has three spatial axes: time, work process, and spatiality. Past working history generates automatically according to user's data analysis. We can get this history as logs of interactions.

Visualization of past working history offers three advantages for the user:

1. The ability to access previous state/process and manipulate data at that state
2. The accessibility to understand the relationship between current state and the procedures taken to achieve it
3. It can act as a work memory support for the user

We propose visualization of past working history because of the three advantages. Our concept of past working history visualization is shown in Figure 1. Data is visualized in a point that a user manipulates the data as a trial. The user can interact with the visualized result and can manipulate it. According to the user's analysis, the data is visualized and past working history is accumulated. The data mining process is visualized like a road. The road represents analysis process and a road branch represents a process branch. Branches of road indicate user's trials histories. When the user takes a bird's eye view, the user can get the whole view of the analysis. The user can access the working history and interact with them.

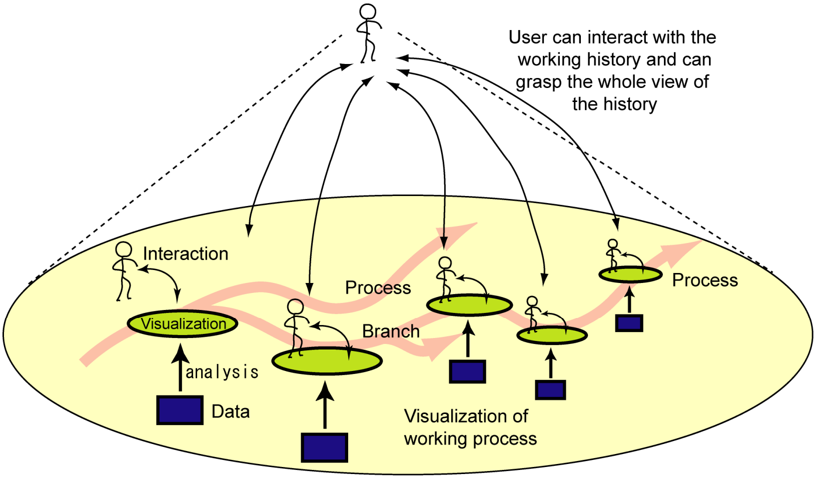
User can interact with the working history and can grasp the whole view of the history
Interaction
Visualization
analysis
Data
Process
Branch
Process
Visualization of working process

Fig. 1. Concept of visualization of past working history. Past working history is visualized like a road and connected each visualized result. Users can grasp the whole view of the history as bird's eye view and interact with them.

To visualize past work history, there are two kinds of visualization methods. One is visualization with three axes written above. The other is visualization without any axes. This visualization method includes footprints as a browser history and markers as memo. In this paper, to visualize past work history efficiently, we have developed three categories for visualization method using the three axes:

1. Past Working History Visualization with Time Axis
2. Past Working History Visualization with Process Axis
3. Past Working History Visualization with Spatial Axis

Time axis is generated automatically. As time naturally flows, the work process is generated by users' analysis. Spatial axis is generated by user's interactions. We use these axes for visualization of past working history. When data has several states and the user do trial and error approach, we define the state position and the history is visualized that are taken one axis as time and the other axis as state. When data has hierarchic structure, we use this structure in the visualization. For example, a user conduct a clustering method manually, data is divided from top to bottom. The clustering process can generate a dendrogram (tree diagram) that indicates data structure. We use this structure as history and add functions on the dendrogram. When the user interacts with visualized objects such as slide-bars that enable the user to set parameters, slide-bars remain as the user set. This slide-bars status is a spatial history. When the user transforms the location of visualized objects, the previous position is a spatial history. We also visualize these kinds of spatial histories.

To visualize the history, various kinds of log data are required in the analysis. Log data should include time, analysis state, parameters, thumbnails, and relationships between multiple states. There are many ways to get log data. For example, by logging application's data parameter, web browser history, logger software which takes logs, OS logs, or other devices such as RFID cards which contains usage logs.

4 Applications and Results

4.1 Applications

To evaluate our visualization method, we applied it to two tasks. First is for web browsing. Second is for visualizing past work history for data mining in genome science.

When we perform a search on the Internet, we put the query into a search engine such as google and yahoo, and check several sites whether we obtained the correct information. We sometimes follow a link to go forward to other sites. The web browsing process has a structure. In this example, the search engine is at the top and several sites are the next step to the search engine, then further sites are the tertiary step. We visualize this structure in order to grasp the web browsing process.

In the field of genome science, researchers are facing huge experimental data. Technology of sequencing and microarray enable us to get large biological data at a time. Using these high-throughput technologies, experimental data are accumulated rapidly and genome researchers have to analyze data and interpret them biologically and medically. Genome data consists of patients, expression levels of genes, copy number of genomes, sequence of genomes, and so on. Genome researchers observe and check the data in whole chromosomes level, or in one chromosomal level, then zoom up into a chromosomal band level. Finally, they could also zoom up to a genome sequence level. In this search process, they want to find abnormal regions from the data and to grasp the relationship between abnormal regions. Consequently they will compare among different patients because they want to learn if the abnormal regions are isolated cases or common among many individuals.

Thus we apply our visualization methods to the field of genome science and visualize their searching and checking data histories to support their analysis.

4.2 Visualization with Time Axis

We applied visualization method using time axis for web browsing history. We took Firefox browser's history and captured the page screenshots using ImageMagic library. Visualization is based on OpenGL.

We visualize the web browser history in two and three dimensional spaces. Z direction in three dimensional visualization is the time axis and Y direction in two dimensional visualization is also time axis. We distribute web site screenshots according to the history in time series. The same domain web site is located in the same position in a space. The result of visualization with time axis for web browser history is shown in Figure 2.

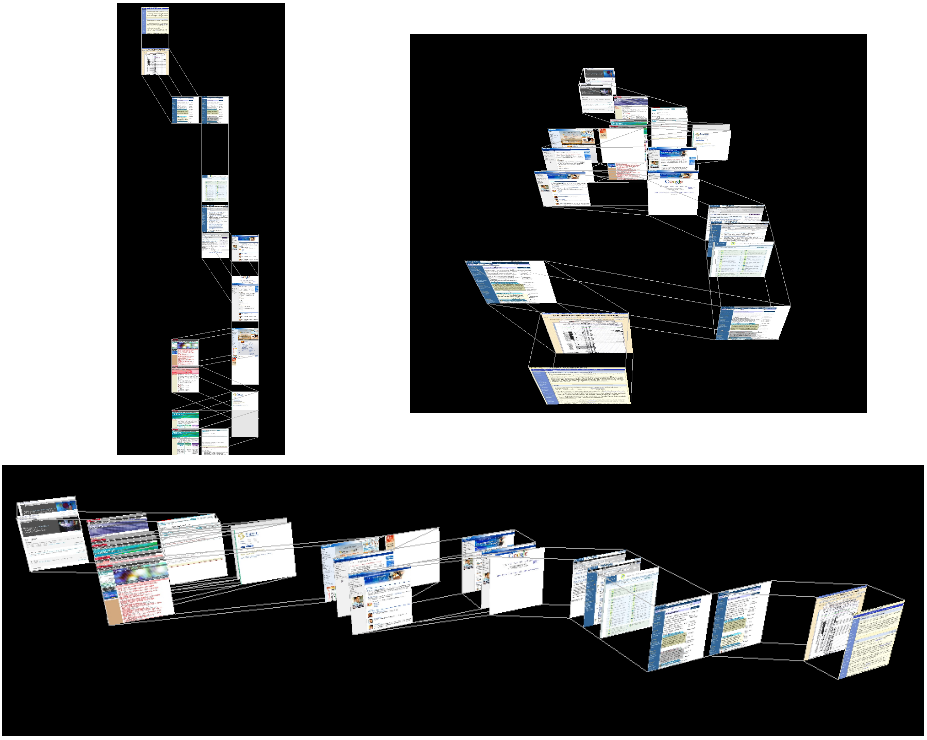

Fig. 2. (upper left) Visualization with time axis in two dimensional spaces. X axis represent domains and y axis represents time. Web thumbnails are presented. (upper right) Visualization with time axis in three dimensional spaces. Domains are distributed in 2 dimensionally and Z axis represents time. (lower) Visualization with time axis in three dimensional spaces. Z axis represents time and X axis represents domains. Web thumbnails are stood using X and Y dimension.

We also visualize genome science application history. Genome science researchers look the data in whole view level and in each chromosome level. We have developed genome copy number viewer for supporting genome researchers. The viewer enables them to present both whole chromosome view and each chromosome view. User can compare patients' data using the viewer. We have obtained logs of this viewer and visualized them at a same time. We took one axis as time and others as the relationship between whole and part views. That is, data structure is distributed two dimensionally. When the user searched for data from chromosome 1 to chromosome Y, the history visualized a pattern. The result of visualization is shown in Figure 3. The result reveals their analysis process that they check the whole view and see the details view, then they go back to whole view and repeated this cycle.

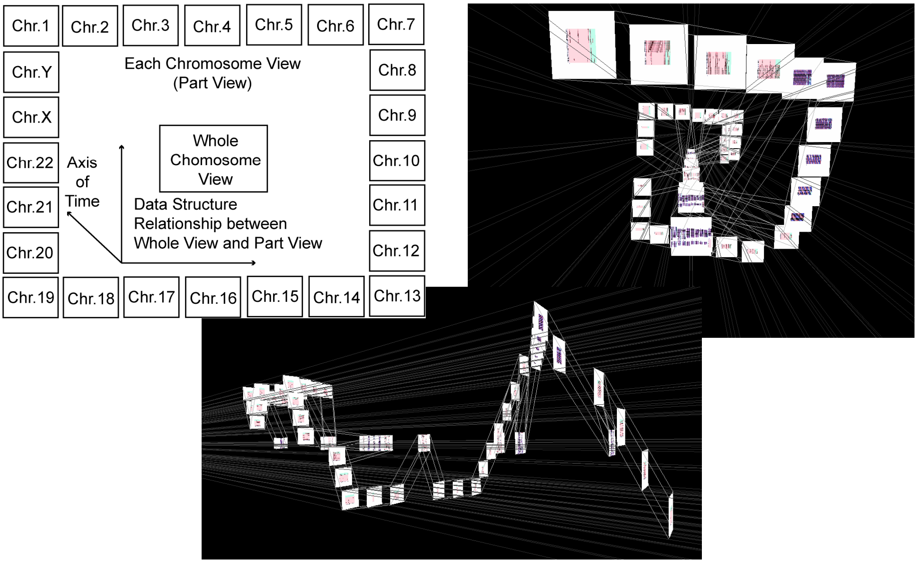

Fig. 3. (upper left) Distribution of data. Data structure is located in two dimensional spaces and the other axis indicates time. Each chromosome view is located in a rectangle. Center of the view is whole chromosome view. (upper right) Visualization with time axis and data structure. There are 24 locations for each chromosome view and 1 center location for whole chromosome view. When a user see and check the data according to chromosome number, each chromosome visualization results are distributed spirally. (lower) The same view from different angle. Time axis is visualized from right to left. Each views which visualized as thumbnails are connected each other, thus a user can grasp transitions of the work history. The user can tell that user checked the data whole view and each view to-and fro.

4.3 Visualization with Process Axis

We applied visualization with process axis for browsing application. We took the interaction history from our original application for browsing genome data.

We have developed the graph viewer for genome data. The viewer is for analyzing genome copy number and genome researchers want to detect abnormal copy number regions. They compare the candidate's abnormal copy number regions that are detected by one patient data with other patients' data to check whether those regions are common. The viewer supports this analysis. It presents graphs that indicate genome copy number for each patient. Graphs are distributed in tile fashion. A user can select a graph to check the detail, the graph will be zoomed in. Then user click the graph again, the view will be zoomed up again. Then, the user can find an abnormal region and click it. The viewer extracts the same region for all patients and distributed in a line. It provides the user a comparison view among patients. Then, the user can know which one is common abnormal region. We got interaction history in this analysis. The viewer outputs the state and the thumbnail at the time when there is any interaction automatically. When the user analyzes the data, the interaction history is accumulated automatically.

The analysis has a hierarchy. There are five steps: whole view, selection of one view, zoom view, the selection of an abnormal region in the view, and comparison among patients. We visualize this process as an axis with a dendrogram (tree structure) that shows relationship between whole and part views. Figure 4 shows the viewer and history visualization result.

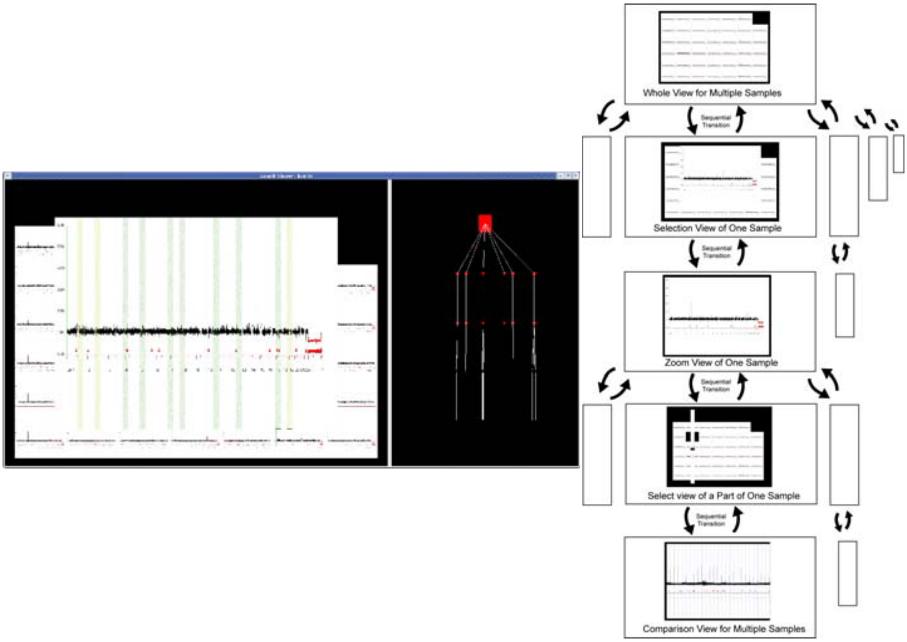
Whole View for Multiple Samples
Sequential Transition
Selection View of One Sample
Sequential Transition
Zoom View of One Sample
Sequential Transition
Select view of a Part of One Sample
Sequential Transition
Comparison View for Multiple Samples

Fig. 4. (left) Distribution with process axis. The left window indicates a viewer and right window indicates visualization of work history. Work history is visualized as a dendrogram which is tree structure. (right) Work process has a hierarchal structure and it can be visualized as a dendrogram. This figure indicates the work process in an image. There are 5 levels. A user can see whole view for multiple samples. Then the user can select one sample. The selected view is zooming up. The user can select one region where the user feels abnormal. The selected region can be compared among multiple samples in the last view. This process is visualized as a dendrogram in right image.

4.4 Visualization with Spatial Axis

When a user interact with virtual objects and move them spatially, we can visualize the interaction history with spatial axis. One example is the slide bar. A user can set parameters by manipulating slide bars. It is implemented with three dimensional positioning sensors. The other example is moving histories. When a user move or fly in the virtual environment, the trajectory indicates history. We visualize the history with spatial in a virtual environments.

5 Summary and Discussion

In this paper, we proposed three visualization methods. We introduced two applications for our visualization methods as results.

The results indicate that history is useful to grasp the process. However, past history visualization has limitation because it depends on the amount of history data. There is still room for improvement to further incorporate and utilize history of interactions.

Acknowledgments. This work is supported in part by Core Project from Microsoft Institute for Japanese Academic Research Collaboration (IJARC).

References

1. Card, S.K., Mackinlay, K., Shneiderman, B.: Readings in Information Visualization: Using Vision to Think (1999)
2. Gershon, N., Eick, S.G., Card, S.K.: Information visualization. In. ACM interactions 5, 9–15 (1998)
3. Nishimura, K., Ishikawa, S., Hirota, K., Aburatani, H., Hirose, M.: Information Visualization of Allelic Copy Number Data for Genome Imbalance Analysis. 11th International Conference on Human - Computer Interaction (HCI International 2005) CD-ROM (2005)