A Visualization Technique for System Logs Realizing Overview and Compression

Aki Hayashi* Takayuki Itoh[†]
Ochanomizu University Ochanomizu University

1 Introduction

A lot of research has been conducted on visualization of time series data recently. Among them, we are focusing on multi dimensional and large scale time series data, for instance system logs such as transactions of credit cards and access logs of Web sites. The transaction data we have contains 160,000 records in 6 months, and each record has 40 attributes, such as credit card ID, shop ID, fraud ID and item code. The access log files of our laboratory's Web site have also 200,000 lines in 6 months and there are 12 attributes such as IP address, URL, status code and so on. The aim of observing system logs as time series data is as follows: from transaction records, discoveries of the tendency of overall transactions, particularly fraudulent transactions, and from access logs, discoveries of the tendency of access, observations of error occurrences, and determination of renewal timing. However, system logs are so massive that to obtain significant knowledge from such data is difficult and often only wasting much time.

This poster presents a visual analytics tool which enables us to effectively observe system logs as time series by visualization technique realizing overview and compression. This system also realizes recommendation of attributes selections which bring us interesting discoveries. To overview the system logs, we assigned time to X-axis and values of each attributes (items) to Y-axis. Then the system depicts the total of grid (corresponding to one time interval of an item) like a heatmap. In addition, the system calculates and shows the degree of recommendation of each attributes to lead effective results.

2 Rerated Work

Some approaches[1][2] have realized attribute recommendation for specific visualization methods including scatterplot, in the context of visualization for general multi-dimensional data. However, most approaches which focus on time-series[3] have not considered recommendation of attributes which support the analysts to obtain effective results. In addition, there are few works which make use of compression of drawing result which has already realized for 1-dimensional time-series data. On the other hand, our method focuses on time-series of multi-dimensional data, and then realizes attribute recommendation based on the significance of the leading result and two kinds of compressions. WireVis[4] is a visual analytic tool for system logs which realizes a set of coordinated visualizations with multiple views. We also scrutinize several discoveries from proposed method using scatterplot[2] assigning non-time-series attributes to X-axis aiming at results by WireVis[4]. Attribute selection for scatterplot[2] is a more complicated problem than our problem; however, using our method together with scatterplot[2], more interesting discoveries can be expected.

3 Visual Analytic Tool for System Logs

3.1 Overview of characteristic of time-series

The proposed method assigns time-series to X-axis, and attributes values (items) to Y-axis, and then depict total of each time and value like a heatmap. It provides choices for the X-axis: days, day of the week, months, and hours. The users can also select all attributes including time-series for Y-axis. If the total is larger, the method depicts the grid by warmer colors. The system shows detailed information of the grid such as names of items and total values by mouse click. The users can also eliminate drawing results by specifying particular values of attributes. We call this functionality "filter".

3.2 Recommendation of attribute selection for Y-axis

The presented technique recommends the preferable attribute which leads interesting results when the user selects the attribute to visualize. Our system calculates one of the following criteria selected by the user. Here, the recommendation degree is calculated as relative value.

(1) The average of the total value of all grids
(2) The average of the maximum value of each item
(3) The maximum value of sum total of each item
(4) The height of the entropy of all grids
(5) The lowness of the entropy of all grids

If value of (1) is high, there should be at least one remarkable grid. The value of (2) is high when there are many remarkable grids. The value of (3) judges whether there are at least one remarkable item. When the value of all grids spread, the value of (4) should be high, and vice versa.

3.3 Compression of the visualization result

It is possible that the system cannot comprehensively draw all items when the overview result (see Section 3.1) has too many items. For example, there appears about 6,000 URLs in our access log files of 6 month. In such case, there are too many horizontal belts and if the user transfers or zooms up the result, the drawn items may change. Thus we propose two kinds of compression.

3.3.1 Compression by sorting algorithm

The system sorts the items by one of the following criteria selected by the user. This time the system calculates the value of following criteria as an absolute value.

(1) The maximum value of each item
(2) The maximum value of the amount of increase of each item
(3) The height of the entropy of each item
(4) The lowness of the entropy of each item

When the value of each item spread, the value of (3) should be high, and vice versa.

2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610 JAPAN
*e-mail: aki@itolab.is.ocha.ac.jp
[†]e-mail: itot@is.ocha.ac.jp

3.3.2 Compression by clustering algorithm

The compression by sorting algorithm (see Section 3.3.1) sometimes lost the interesting items for instance there are some items which has same periodic behaviors but the value of each grids are not so high. Therefore, we propose another compression functionality applying k-means clustering algorithm. Our system makes 30 clusters and selects a representative item of each cluster to compress the result. In addition, the system can also draw the all items that belong to particular cluster selected by the user.

4 Examples

4.1 Analysis of the tendency of overall transaction

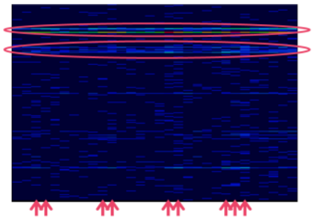

Figure 1. Tendency analysis from transaction records of 11/2007

Figure 1(left) shows the overview result (X: days, Y: shop ID). We can understand that there are several shops which have continuously many transactions surrounded by red circles. We can also understand that on weekends or national holidays (indicated by red arrows) there are periodic increases of transactions among several shops. In this year there was successive holiday from 23/11 to 25/11. Figure 1(right) shows the items belong to certain cluster after applying the clustering. We can observe that most shops which have large transactions on the last day of the successive holiday have also large transactions on another 2 days of the holiday or other weekends.

4.2 Analysis of the fraudulent transactions behavior

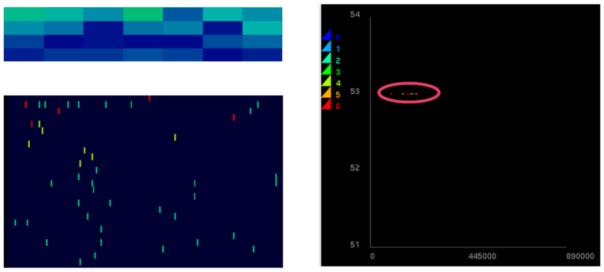

Figure 2. Tendency analysis from fraudulent transactions of 07/2007-12/2007

Figure 2(upper-left) shows the part of visualization result (X: day of the week (Sun.-Sat.), Y: item cord). Although merchandise sold in foreign countries (first line) has continuously large fraudulent transactions, electrical appliances (second line) and train tickets (third line) has more fraudulent transactions on weekends, and mail orders (fourth line) has more on week days. Figure 2(lower-left) eliminates item cord to electrical appliances (X: days, Y: shop ID recommended by (3)) and apply the compression sorted by (1). There are continuous fraudulent transactions at the shop drawn on second line. By changing Y-axis to card ID, the card ID of those transactions proved to be non-identical. Consequently, we scrutinized this shop using scatterplot[2] (X: amount, Y: fraud ID, color: day of the week). It is proved that on the weekend, the transactions around 1,000 dollars using forged credit card are concentrated. This shop has the possibility to be the target of forged cards.

4.3 Observation of access and error, and suggestion of renewal timing

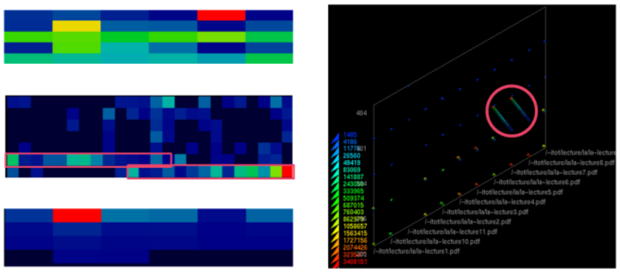

Figure 3. Tendency analysis of access logs (04/2011-09/2011)

Figure 3(upper-left) shows the example (X: months, Y: URL) applied the compression sorted by (1). The original number of items was about 6,000 and the readability of the result was remarkably improved. Of course, the top and index pages are found from this example because they must be frequently visited. In addition, 7th and 8th lecture's hand-outs of one of the professor's lectures are also found from same result. From Figure 3(center-left) we observed the timing of access of 7th lecture (X: hours (0-24), Y: day of the week (lower: Sun., upper: Sat.)). To prepare the lecture held on Monday morning, many students might access this page from Sunday evening to early morning on Mondays, especially 23:00 on Sundays. Thus, the professor should renewal the page before that timing. Consequently, we selected X: month, Y: status code leading from recommendation (4) shown in Figure 3(lower-left). Here we found that status code 206 (partially completed) occurred more frequently than 200 (completed) even though 200 had to be more general. From this result, we can expect that this 206 might be one reason of such large access. Therefore we scrutinized all hand-outs of this lecture using scatterplot[2] shown in Figure 3(right) (X: URL, Y: status code, color: transferred data amount, height: access number). As we expected, there were many accesses whose status code was 206 only for the 7th and 8th hand-outs of this lecture. We also found that the IP addresses which cause this error were diverse, thus we could discovered the possibility that many student might had difficulty in getting the hand-outs of 7th and 8th lectures.

5 Conclusion and Future Work

We have presented the visualization technique and the visual analytics tool to effectively overview and compress the system logs as large scale multi dimensional time series. As future work, we are discussing about improvement of the color settings and indicating the patterns repeatedly occur across the items.

Acknowledgements

The authors would like to thank Intelligent Wave Inc. providing the transaction records of credit cards.

References

[1] A. Tatu et al., *Combining Automated Analysis and Visualization Techniques for Effective Exploration of High-dimensional Data*, IEEE Symposium on Visual Analytics Science and Technology, pp. 59-66, 2009

[2] C. Sakoda et al., *Visualization for Assisting Rule Definition Tasks of Credit Card Fraud Detection Systems*, IIEEJ Image Electronics and Visual Computing Workshop, 2010.

[3] W. Aigner et al., *Visualizing time-oriented data - A systematic view*, Computers and Graphics, Vol. 31, No. 3, pp. 401-409, 2007.

[4] R. Chang et al., *WireVis: Visualization of Categorical, Time-Varying Data From Financial Transactions*, IEEE Symposium on Visual Analytic Science and Technology, pp. 155-162, 2007