An Associate-Rule-Aware Multidimensional Data Visualization Technique and Its application to Painting Image Collections

Ayaka Kaneko
Grad. School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
ayaka@itolab.is.ocha.ac.jp

Akiko Komatsu
Grad. School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
akiko@itolab.is.ocha.ac.jp

Takayuki Itoh
Grad. School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
itot@is.ocha.ac.jp

Florence Ying Wang
CSIRO
Sydney, Australia
Florence.Wang@csiro.au

***Abstract*—This paper presents a visualization technique for multidimensional datasets containing real and categorical variables. Supposing multidimensional datasets containing real and categorical values, this technique displays a set of axes corresponding to the dimensions of real values. The technique evenly divides the axes into several ranges and displays component bar charts there. It brightly draws the component bar charts if association rules are applied at the corresponding ranges of the dimensions of real values. As a result, this technique highlights association rules so that users can discover important relationships between real and categorical variables in multidimensional datasets. This paper introduces an application of the presented technique to painting image collections. This application visualizes image features and categorical information of painting images and provides a user interface to browse the painting images associated with the multidimensional values. This paper also introduces user evaluation results of the user interfaces for painting image collections.**

***Index Terms*—Multidimensional data, Association rule, Image feature.**

I. Introduction

Multidimensional data visualization has been a well-studied research topic. Multidimensional data which visualization techniques deal with may generally contain numeric, ordinal, and categorical values [1]. We have various multidimensional datasets containing such complex types of variables in our daily life. For example, transactions of consumer shops contain numeric values such as number of visitors and amount of revenue, ordinal values such as date and time, and categorical values such as name of shops. Association rule mining techniques [2] have been well-studied to discover interesting relationships among such variables. We expect multidimensional data visualization techniques featuring association rules would help users to easily discover and understand important relationships among the variables of multidimensional datasets.

This paper presents a multidimensional data visualization technique which emphatically represents association rules. This technique displays a set of parallel axes corresponding numeric variables of an input multidimensional dataset. The technique also displays small component bar charts for intervals of the axes and highlights the component bar charts corresponding to the intervals which one or more association rules are applied.

This paper also presents an application of the visualization technique as a browsing application of painting images. This application supposes a set of painting images, calculates image features as numeric values, associates textual attributes of the images as categorical value, and visualizes the association rules between the numeric and categorical values. The application also provides user interfaces to support users to exploratory discover users' favorite painting images from the highlighted association rules.

II. Related Work

A. *Multidimensional data visualization*

Scatterplot matrix (SPM) and parallel coordinate plots (PCP) are especially well-studied multidimensional data visualization techniques. SPM is preferable when users want to observe correlations between arbitrary pairs of dimensions; however, it requires large display spaces when an input dataset contain large number of dimensions. PCP is preferable to represent distributions of numeric values of a set of dimensions in a small display space; however, it often causes visual cluttering when large number of individuals are contained in an input dataset. Also, it is difficult to discover correlations between non-adjacent pairs of dimensions.

B. Association rule and visualization

Itoh et al. [3] presented a multidimensional data visualization technique which extracts a set of low-dimensional spaces and represents as a set of small PCPs. The technique provides two methods for low-dimensional space extraction methods supposing that input multidimensional datasets contain numeric and categorical variables. One of the methods extracts groups of dimensions which have strong correlations. The other extracts groups of dimensions which one or more association rules are applied with user-specified categorical variables.

In our survey, the above technique is one of the visualization techniques which aggressively applied association rules. However, it is still difficult to clearly discover which ranges of numeric variables satisfy the association rules while observing the visualization results of the above technique. To solve the problem, this paper presents a visualization technique which highlights the ranges of numeric variables which satisfy the association rules. Though there have been several visualization techniques addressing the representation of association rules [4], [5]; however, there have been few studies which visualize confidence and support values of association rules while simultaneously representing numeric distributions of input multidimensional datasets.

C. Image browser and visualization

This paper presents a painting image browser which assists users to explore the collection of painting images with their image features and labels. In other words, the image browser supposes numeric and categorical variables are associated with each of the painting images.

Image browser is a recent active research topic. PhotoMesa [6] and CAT [7] are typical image browsers for hierarchically structured sets of image collections. These techniques apply space-filling algorithms which place a set of images in rectangular subregions to represent the hierarchy of the images.

There have also been several techniques for browsing unstructured sets of image collections. SIB (Semantic Image Browser) [8] applies a dimension reduction scheme while D-FLIP [9] applied a force-directed layout to place a set of images into the screen space.

ImageCube [10] is an image browser which supposes multidimensional vectors are associated with each of input images. ImageCube displays a set of images assigning a pair of variables onto horizontal and vertical axes so that users can observe the relationships between appearances of images and values of the multi-dimensional vectors.

III. Associate-Rule-Aware Multidimensional Data Visualization Technique

The visualization technique presented in this paper highlights the ranges of numeric variables where one or more association rules are discovered. Also, the technique visualizes the ratio of user-specified categorical values, as well as confidence and support values at each of the ranges. The

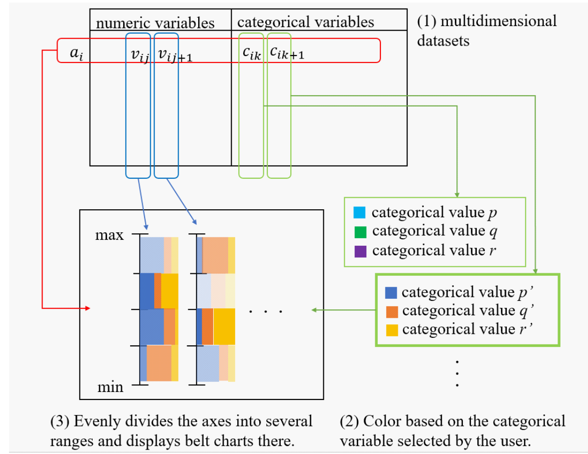

Fig. 1. Processing flow of the presented technique.

section describes the data structure and processing flow of the presented technique.

A. Data structure and processing flow

We suppose a multidimensional dataset D has n individuals, and an individual a_i has m_v real variables and m_c categorical variables. We formalize the dataset as follows:

$$D = \{a_1, a_2, ..., a_n\}$$
$$a_i = \{v_{i1}, v_{i2}, ..., v_{im_v}, c_{i1}, c_{i2}, ..., c_{im_c}\} \quad (1)$$

Here, v_{ij} is the j-th real variable of the i-th individual, and c_{ik} is the k-th categorical variable of the i-th individual.

Fig. 1 illustrates the processing flow of the presented technique. The technique assigns colors to each value of the user-specified categorical variable as shown in Fig. 1(2). Then, the technique displays vertical axes corresponding to real variables, divides each of the axes into k ranges, and displays small component bar charts at each of the ranges as shown in Fig. 1(3). The component bar charts depict the ratios of categorical values in the corresponding range. The number of individuals corresponding to a range is depicted by the intensity of the corresponding component bar chart. This representation of component bar charts tightly relates to the representation of association rules as described later.

B. Association rule for visualization

This technique extracts ranges of numeric variables where satisfy one or more association rules with categorical values. Here, let A as a specific range of a numeric variable, and B as a specific categorical value. The definition of an association rule in this paper is a combination of A and B which satisfies $A \rightarrow B$ or $B \rightarrow A$. It denotes that there are many individuals in the range A which has the categorical value B. Our implementation calculates the support P_{sup} and the confidence P_{con} for arbitrary pair of A and B by the following equations, and extracts A and B if both P_{sup} and P_{con} are larger than user-specified thresholds.

$$P_{sup} = P(A, B)$$
$$P_{con} = (B|A) \tag{2}$$

This definition is often satisfied at the ranges where corresponding component bar charts have large portions of specific categorical values and are brightly colored. In other words, our representation of component bar charts depicts how the ranges of numeric variables satisfy association rules.

The user interface we developed for this technique features slider bars to adjust the thresholds of support and confidence values. We can control the number of ranges satisfying the association rules by operating the slider bars.

IV. Application to the Image Collection

This section presents an application of the presented visualization technique for browsing of painting image collections. This application treats annotations and image feature values as numeric and categorical variables, and assists users to explore the image collections while highlighting association rules as clues for users' exploration.

A. Image features

Design of image feature has been an active research topic since CBIR (Content Based Image Retrieval) was expected as a powerful approach for image retrieval. A survey paper by Zhou et al. [11] introduces algorithms and image features for CBIR; where image features include global features based on color, shape, texture, and composition, as well as local features such as SIFT. Recent image processing communities often call such traditional image features as "hand-crafted features." On the contrary, recent image retrieval techniques apply learning-based features as an abstract representation of image features constructed by deep neural networks.

This application applied hand-crafted features because we preferred explanative feature values for our application.

Table I shows the list of image features. We applied global features of colors, edges and compositions proposed by Wang et al. [12], as well as image features calculated from DoG filter, Gabor filter, and local color histogram. Color features implemented in this study include color temperature score and color weight score. Here, our implementation decomposes an image into pre-defined number of local colors and calculates the temperature and weight for each local color. The number of local colors was three in our study. Line feature is calculated from the amount of edges extracted by Hough function. Composition feature is calculated from the saliency model presented by Itti and Koch [13].

Some of the image features including color temperature score, color weight score, and line are visualized as numeric variables because it is easy to understand the relationship between the appearance of painting images and these values. On the other hand, we apply a clustering algorithm with other image features including composition, Gabor filter, DoG filter, and local histogram, and treat cluster IDs assigned to painting images as categorical variables. Our implementation firstly applies principal component analysis (PCA) to these features, and then clusters by k-means algorithm. We measured the quality of clustering results by applying Davies-Bouldin Index [14] to specify the optimal number of clusters. As a result, we constructed seven clusters with composition, six clusters with Gabor filter, seven clusters with DoG filter, and two clusters with the local histogram.

TABLE I
Image feature applied in this study.

Feature value	Description	Dimensions
Color temperature	Visual temperature of color of three segments	3
Color weight	Visual weight of color of three segments	3
Line	Straight line ratio, mean slope, mean length, standard deviation of slopes	4
Composition	Mean saliency for each of the nine image regions divided by "Rule of thirds"	9
Gabor filter	Multi-directional high frequency component	40
DoG filter	Differential feature values independent of direction	6
Local histogram	Local gray-scale histogram	90

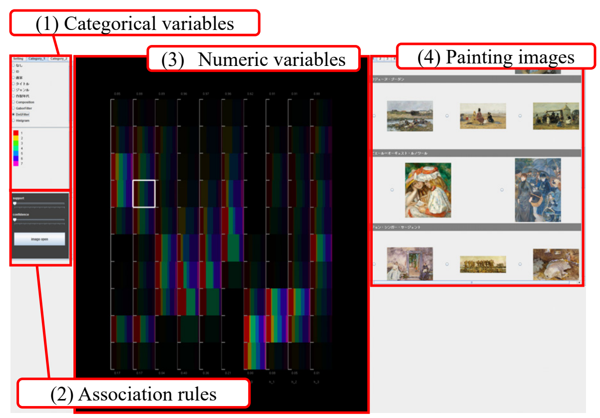

Fig. 2. Snapshot of our application consisting of four user interface components.

B. User interface

Fig. 2 shows the user interface of our application. The user interface consists of the following four components:

- Category selection panel (CSP) (Fig. 2(1)),
- Association rule control panel (ARCP) (Fig. 2(2)),
- Image feature visualization area (IFVA) (Fig. 2(3)), and
- Painting image display panel (PIDP) (Fig. 2(4)).

IFVA displays a set of component bar charts to represent image features treated as numeric variables. The component bar charts depict the distribution of categorical values of a variable selected on CSP. It also displays the categorical values and their corresponding colors used in the component bar charts. Fig. 3 shows an example of CSP. Users can select a categorical variable by pressing a radio button in the upper part

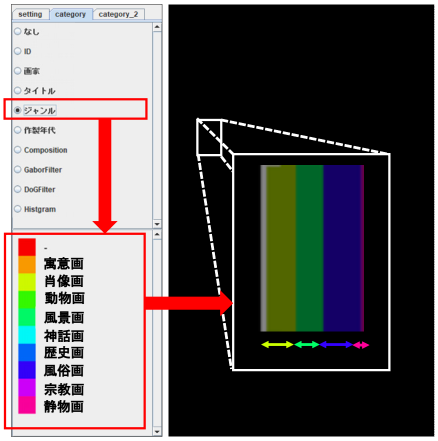

Fig. 3. Categorical variable selection by radio buttons. The color palette displays a list of colors and corresponding categorical values. Component bar charts visualize image feature values.

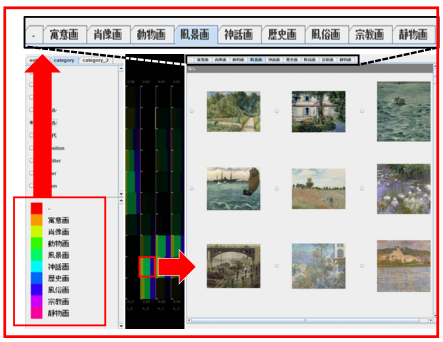

Fig. 4. Corresponding painting images are displayed according to the click operations of component bar charts.

of the panel, Categorical values and corresponding colors are then displayed in the lower part of the panel, and component bar charts in IFVA are then updated.

When a user clicks a particular component bar chart, corresponding painting images are displayed in PIDP as shown in Fig. 4. This panel generates multiple tabs corresponding to categorical values of the user-specified categorical variable and displays painting images divided according to their categorical values. Also, our implementation sorts the displayed images according to another categorical variable selected by users on CSP. This section calls the former categorical variable "coloring category" and the latter variable "sorting category." When a user clicks a painting image, IFVA highlights the corresponding component bar charts, as shown in Fig. 5. This highlight can be a clue for users to look for similar images.

IFVA can skip to draw component bar charts corresponding to the ranges which satisfy no association rules. Also, users can adjust the thresholds of confidence and support values on ARCP. Fig. 6 shows an example of the operation of the thresholds. Ranges containing large number of images are highlighted when we adjust the threshold of support value larger. Or, ranges containing large number of specific category of images are highlighted when we adjust the threshold of

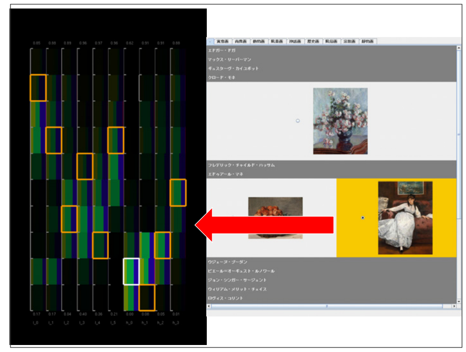

Fig. 5. Subranges corresponding to the user-selected painting image are highlighted by orange borderlines.

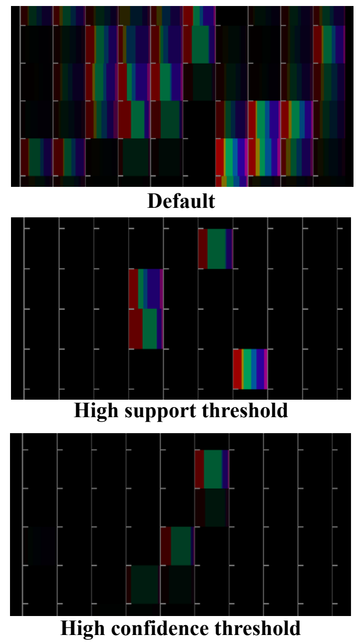

Fig. 6. Threshold adjustment of association rule by using slider bars.

confidence value larger.

C. Image collection

We downloaded 858 painting images which are annotated as "impressionism" from a public domain Web site [1] and constructed a dataset from these images. The dataset contains image files, image feature values, clustering results of some of image feature values, and annotations such as artist names, year of the production, and genre of the painting. Following is the list of variables (See Table I) and annotations of the dataset:

- Categorical variables
  - Annotations (Artist name, Year of production, Genre of the painting)
  - Clusters (Gabor filter, DoG filter, Local histogram, Composition)

[1] http://free-artworks.gatag.net/

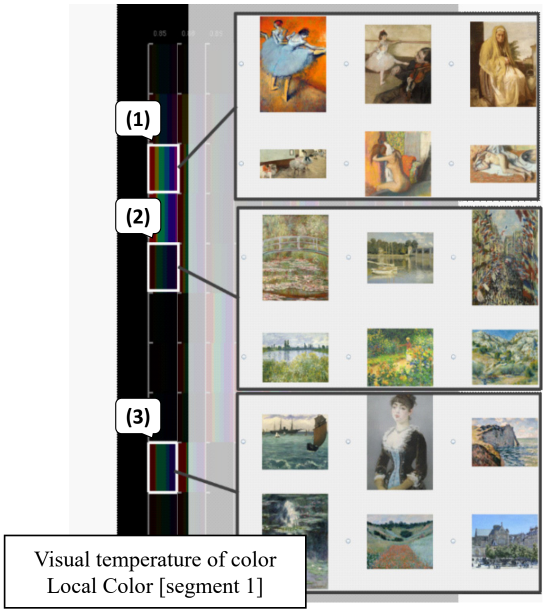

Fig. 7. Visualization of color temperature score of the first local colors. Reddish, greenish, and blueish colors correspond to (1), (2), and (3) in this snapshot.

- Real variables
  - Color temperature score, Color weight score, Line score

We aim to assist users to easily explore image collection and discover their favorite ones under the following conditions:

- Discover their favorite painting images which they do not know the names of paintings and their artists.
- Discover unknown paintings which are similar to their favorite famous painting.

We suppose annotations and image features can be great clues for users to discover their favorite paintings. The application visualizes such information to assist users to understand the attributes of their favorite paintings and make this knowledge clues of their exploration of image collections.

D. Example

This section introduces examples of visualization and use-case scenarios. Here, we used the presented application supposing the situation that we would like to look for our favorite painting images without any appropriate query words or query images. We applied an image dataset described in the previous section in this experiment.

We observed IFVA at first. The left end axis displayed there corresponds to the color temperature score of the first local color. Fig. 7 shows three types of painting images. The range (1) contained reddish and warm painting images. Many greenish images belonged to the range (2). We discovered our favorite blueish images in the range (3).

Here, we observed features of blueish images belonged to the range (3). We selected DoG filter as the coloring category at that time. Then, the component bar charts in

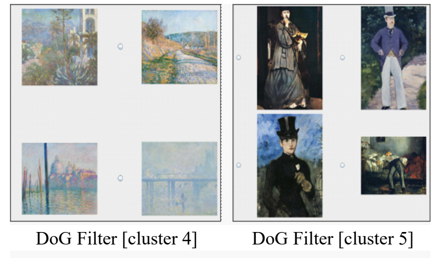

Fig. 8. Painting images clustered based on DoG filter feature values.

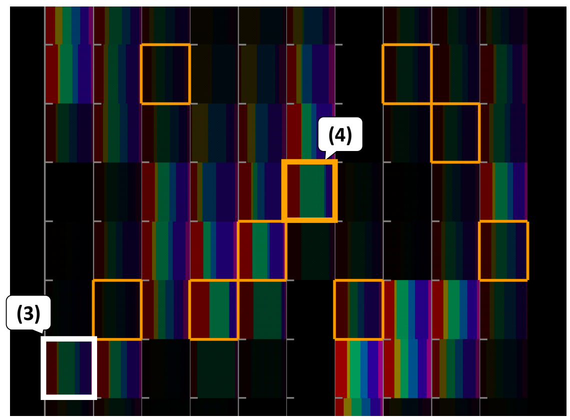

Fig. 9. Visual discovery of association rules based on distribution and intensity of component bar charts.

IFVA were painted in seven colors corresponding to the seven clusters generated with feature values of DoG filter. Also, PIDP generates seven tabs corresponding to the seven clusters and displays images belonging to each of the clusters. Fig. 8 shows examples of images in clusters 4 and 5. Images in cluster 4 have pale blue backgrounds and soft edges. On the contrary, images in cluster 5 have vivid colors and clear edges. Or, we can browse images divided by other categories selected as the sorting category on CSP instead of DoG filter.

Then, we specified a favorite painting image and looked for other images which have similar feature values. When we click the favorite image, component bar charts corresponding to the ranges of numeric variables which the clicked image belongs to are highlighted by orange borderlines as shown in Fig. 9. This visualization shows the range (4) in Fig. 9 has many images belonging to cluster 4 depicted in green. Also, this range has a relatively large number of images because the corresponding component bar chart is brightly rendered. It suggests that cluster 4 will satisfy an association rule with this range. Here, the color weight score of the third color is assigned to the axis of the range (4). It denotes that the range (4) has images painted in light colors. The appearance of IFVA as mentioned above will be clues to discover painting images similar to users' favorite images.

This application enables users to narrow down the exploration of image collections with the features and category

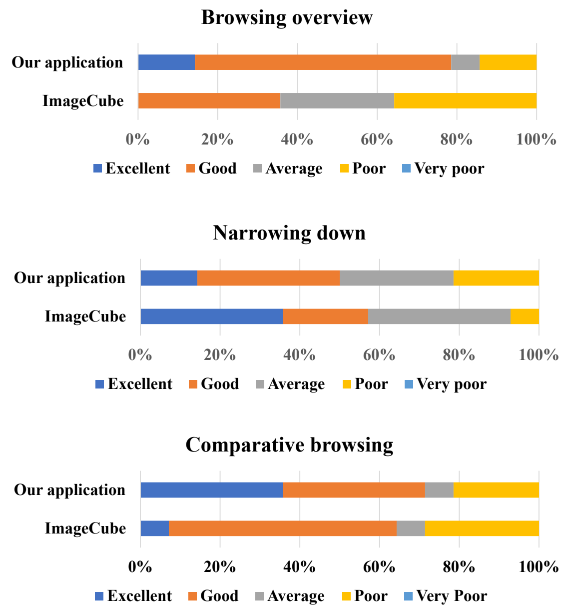

Fig. 10. User evaluation results of the application for painting image collections.

information of the images. Users can discover their favorite images by comparatively looking at images which have similar features or belong to the same categories. Also, association rules visualized by this application can be clues to browse images which have similarities with users' interested images.

E. User evaluation

We had a user experiment of the application while comparing with ImageCube [10]. We invited 14 participants in our campus and conducted a task to look for their favorite images using the presented application and ImageCube. All participants were female university students in twenties because we belong to women's university. After this task, we asked participants to answer the easiness of "browsing overview," "narrowing down," and "comparative browsing" in the five-point Likert scale.

Fig. 10 shows the result of evaluation by the participants. The presented technique archived better results on easiness of browsing overview and comparative browsing. On the other hand, some participants supported ImageCube because it displays images into a 2D space where two of numeric variables are assigned to two axes, and therefore it is easy to narrow down users' interests. On the other hand, the display of numeric variables and image collections are separated in our application. It was not satisfactory for some of the participants while they were narrowing down their interests. We would like to improve our application based on these feedbacks.

V. Conclusion

This paper presented a multidimensional data visualization technique which highlights ranges of numeric variables which satisfy association rules. The paper also presented our application of the visualization technique on painting image browser. We had a comparative experiment with ImageCube [10] and concluded that the application presented in this paper archived better evaluations on overview and comparative browsing.

We need to extend the representation for datasets containing a large number of numeric variables. The dataset introduced in Section 4 was easy to represent by our technique because it included just ten numeric variables. We may need to additionally implement a dimension selection algorithm to visualize only meaningful or important numeric variables in a limited screen space in case a large number of numeric variable is contained in an input dataset.

We also would like to test our technique with larger datasets. The image collection introduced in Section 4 was not large-scale but just contained 858 images limited to paintings by impressionists. We could not discover sufficient variety of association rules with our experiment because the number of painting images is not sufficiently large or categories of the images were unbalanced. We would like to test our application again after constructing more massive datasets containing more variety of painting images.

References

[1] R. Mazza, Introduction to Information Visualization, Springer, ISBN:978-1-84800-218-0, 2009.

[2] T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama, Mining Optimized Association Rules for Numeric Attributes, ACM Symposium on Principles of Database Systems (PODS96), 182-191, 1996

[3] T. Itoh, A. Kumar, K. Klein, J. Kim, High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots, Journal of Visual Languages and Computing, 43, 1-13, 2017.

[4] G. Liu, A. Suchitra, H. Zhang, M. Feng, S.-K. Ng, L. Wong, AssocExplorer: An Association Rule Visualization System for Exploratory Data Analysis, 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1536-1539, 2012.

[5] L. A. F. Fernandes, A. C. B. Garcia, Association Rule Visualization and Pruning through Response-Style Data Organization and Clustering, Advances in Artificial Intelligence (IBERAMIA 2012), 71-80, 2012.

[6] B. B. Bederson, PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps, In Proceedings of the 14th annual ACM symposium on User interface software and technology, 71-80, 2001.

[7] A. Gomi, R. Miyazaki, T. Itoh and J. Li., CAT: A Hierarchical Image Browser Using a Rectangle Packing Technique, 12th International Conference on Information Visualisation (IV08), 82-87, 2008.

[8] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky and M. Ward, Semantic Image Browser: Bridging Information Visualization with Automated Intelligent Image Analysis, IEEE Visual Analytics in Science and Technology, 191-198, 2006.

[9] Y. Kitamura, C. T. Vi, G. Liu, K. Takashima, Y. Itoh and S. Subramanian, D-FLIP: dynamic and flexible interactive PhotoShow, SIGGRAPH Asia 2013 Emerging Technologies Article No. 6, 2013.

[10] Y. Zheng, A. Gomi, T. Itoh, ImageCube: A Browser for Image Collections Associated with Multi-Dimensional Datasets, 15th International Conference on Information Visualisation (IV2011), 12-17, 2011.

[11] W. Zhou, H. Li, Q. Tian, Recent Advance in Content-based Image Retrieval: A Literature Survey, arXiv preprint arXiv:1706.06064, 2017.

[12] Y. Wang, On the Visual Similarity Analysis and Visualization of Art Image Data, Ph. D. thesis, University of Sydney, 2014.

[13] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 1254-1259, 1998.

[14] D. L. Davies, D. W. Bouldin, A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224-227, 1979.