Paper:

Proposal of Treemap-based Cluster Visualization and Its Application to News Article Data

Yasufumi Takama, Yuna Tanaka, Hiroki Shibata

Tokyo Metropolitan University
E-mail: ytakama@tmu.ac.jp

Abstract. This paper proposes Treemap-based visualization for supporting cluster analysis of news article dataset. It is important to grasp data distribution in a target dataset for such tasks as machine learning and cluster analysis. When dealing with multi-dimensional data such as statistical data and document set, dimensionality reduction algorithms are usually applied to project original data to lower-dimensional space. However, dimensionality reduction tends to lose the characteristics of data in the original space. In particular, the border between different data groups could not be represented correctly on lower-dimensional space. To overcome this problem, the proposed visualization method applies Fuzzy c-Means to target data and visualizes the result on the basis of the highest membership values with Treemap. The membership values to the second closest clusters are also visualized, which is expected to be useful for identifying instances around the border between different clusters, as well as the relation between different clusters. A prototype interface is implemented to handle news article dataset, of which the effectiveness is investigated with a user experiment.

Keywords: Visualization, visual analytics, cluster analysis, treemap

1. Introduction

This paper proposes Treemap-based visualization for analyzing the clustering result of multi-dimensional data. When using machine learning and clustering, it is important to understand a target dataset such as its distribution in high-dimensional space. A common approach when visualizing multi-dimensional data is to apply dimensionality reduction methods[9] to target data for projecting it into lower-dimensional (usually 2 or 3D) space. Although this approach improves the visibility of data distribution, it tends to lose the properties the target data has in original high-dimensional space. In particular, the border between different data groups could not be represented correctly on the lower-dimensional space.

As one of the solutions to the problem, this paper visualizes multi-dimensional data without using dimensionality reduction. Instead, the proposed method obtains information about data distribution by applying Fuzzy c-Means[4], and use it for visualization. Different from crisp clustering such as k-Means and AHC (agglomerative hierarchical clustering), soft clustering like Fuzzy c-Means calculates membership values for an instance to all clusters. Therefore, by visualizing not only the relation of each instance to the closest cluster with the highest membership value, but also its relation to the second closest cluster, it is expected that users can identify instances existing around the border between different clusters. It is also expected to be useful for investigating the relationships among different clusters.

We think the proposed visualization method can be applied to different types of multi-dimensional data including statistical data and text data. In this paper, a prototype interface is designed for visualizing text data: a set of news articles written in Japanese. Its effectiveness is investigated with a user experiment.

2. related Work

2.1. Fuzzy c-Means

Different from ordinary crisp clustering, in which each instance belongs to only one cluster, Fuzzy c-Means allows each instance to belongs to multiple clusters with different degree. The degree of an instance being the member of a cluster is called a membership value, of which a range is $[0,1]$. Let an instance $x_k \in X$ (X is a dataset), the membership value of x_k to i-th cluster ($i = 0,\ldots,N_c-1$), μ_{ki}, satisfies the following condition.

$$\sum_{i=0}^{N_c-1} \mu_{ki} = 1, \forall x_k \in X. \quad (1)$$

Membership values are obtained by minimizing the following objective function, where $v_i (\in V)$ is the centroid of i-th cluster, $|\cdot|_2$ is L2-norm, and m is a hyperparameter controlling the fuzziness of the cluster assignment.

$$J(X,V) = \sum_{v_i \in V} \sum_{x_k \in X} (\mu_{ki})^m |x_k - v_i|_2. \quad (2)$$

Zhang et al. have proposed KFCM (Kernel Fuzzy c-Means algorithm)[11], which extends Fuzzy c-Means by adopting a kernel-induced metric instead of Euclidean distance. They reported that KFCM is robust to noise and outliers.

As examples of the application of Fuzzy c-Means, Matsui et al. applied it to analyze the water quality of a lake[7]. It was used to classify the degree of pollution from remote sensing data including noise. Akimoto et al. applied Fuzzy c-Means to detect a human from depth images obtained from the RGB-D sensor[1]. Fuzzy c-Means is used in two ways: a person detection from shape features, and a specific person detection using Fuzzy color histogram.

Visual analytics using Fuzzy c-Means has been studied. Zolkepli et al. have proposed a visualization method for bibliographic big data, in which Fuzzy c-Means is used combined with Newman-Girvan clustering[12]. They claimed fuzzy analysis is expected to offer deeper insights into big data compared with applying crisp clustering. Sherkat et al. have proposed a visual analytics for interactive clustering, in which Fuzzy c-Means is used for term clustering[10].

2.2. Treemap

Treemap[6] is one of common visualization methods for hierarchical structure such as the folder structure of the storage of computers, and the result of hierarchical clustering. Compared with a node-link diagram, which is another common visualization method for hierarchical structure, it can use screen space in an efficient manner.

Usually, Treemap is drawn within a rectangular region. The whole region corresponds to the root of a hierarchical structure, and a hierarchical structure is drawn by recursively dividing a rectangle region of a parent node into sub-regions for children. The size of a rectangle is determined on the basis of the size or importance of its corresponding node. The most famous and basic division method is slice&dice, which slices a parent region into several sub-regions (children) vertically or horizontally, and changes the direction of slicing (dice operation) before further slicing the obtained sub-regions. The operation of dividing a region into sub-regions is called tiling. Alternative Treemaps employing different tiling algorithms from slicing have been proposed. Squarified Treemap[5] divides a region so that the aspect ratio of regions can be close to 1.0. Although it can improve the visibility, the layout tends to change drastically when the size of nodes change dynamically. The Strip algorithm[2] divides a region in one direction like slicing, but it "stacks" multiple layers in a region. Fig. 1 compares (a) slicing, (b) squarified, and (c) strip tiling algorithms for the same structure.

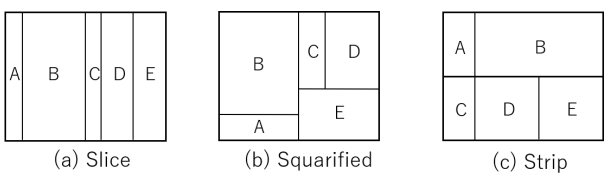

Fig. 1. Comparison of 3 tiling algorithms

As another extension, Voronoi Treemap[3] does not divide a region into a rectangular shape, but into regions of arbitrary shape including polygons.

3. Treemap-based Visualization of Clusters

The proposed method does not visualize data distribution spatially on low-dimensional space. Instead, it visualizes the information about groups (clusters). That is, users do not grasp data distribution on the basis of spatial relationship between instances, but membership values assigned to instances.

As mentioned in Sec. 1, target data in this paper is text (Japanese news articles), to which Fuzzy c-Means is applied. This section describes such a preprocessing for target text data, and visualization method, respectively.

3.1. Preprocessing

A part-of-speech and morphological analyzer MeCab[1] is applied to the title and main text of each news article to extract nouns, adjectives, verbs, and adjective verbs except pronouns, numbers, personal names, and non-independent words. Each article is converted to 100-dimensional vectors by applying Do2Vec[8] of gensim[2] to the extracted words. In addition to that, each article has additional information such as category (business, sports, etc.), publication date, and length (number of characters).

Fuzzy c-Means is applied to the set of news articles: this paper uses scikit-fuzzy[3] with N_c as 5, m as 1.1, stopping criterion (error) as 0.0001, and the maximum number of iteration (maxiter) as 10000.

The clustering result is organized as 2-layer structure: the first layer ($c_i^{rt}, i \in \{0,\ldots,N_c-1\}$) corresponds to crisp clustering result, in which each instance belongs to a cluster with its highest membership value. In the second layer (c_j^i) of c_i^{rt}, each instance in c_i^{rt} belongs to a cluster with its second highest membership value.

$$c_i^{rt} = \{d | c_1(d) = i\}, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

$$c_j^i = \{d | c_1(d) = i \wedge c_2(d) = j\}, \quad \ldots\ldots\ldots \quad (4)$$

where d is an instance (new article), $c_1(d)$ and $c_2(d)$ are cluster labels for those to which d has the highest and the second-highest membership values, respectively.

4. Visualization

Fig. 2 shows the screenshot of the prototype interface, which consists of the following views. It was implemented with HTML5, CSS3, and JavaScript. Treemap and Chord Diagram were implemented with D3.js.

- Treemap view

1. http://taku910.github.io/mecab/
2. https://radimrehurek.com/gensim/models/doc2vec.html
3. https://pythonhosted.org/scikit-fuzzy/

- Cluster information view
- Article information view
- Cluster relation view

The Treemap view occupies the majority of the screen, which shows the result of Fuzzy c-Means. Each cell corresponds to an article, of which color represents cluster labels $c_1(d)$.

The Treemap view has 3 layers: the first layer visualizes all articles, of which the layout is determined by c_i^{rt} as shown in the top of Fig. 3. When a user clicks an arbitrary cell, a zooming operation is applied and only the articles assigned to the same cluster as it are visualized as the second layer. In the second layer, the layout of cells (articles) is determined by c_j^i. When an arbitrary cell is clicked in the second layer, only articles assigned to the same c_j^i as the clicked cell are displayed in the third layer. In the case of Fig. 3, a user clicks a cell assigned to 0-th cluster in the first layer, and articles in c_0^{rt} are displayed in the second layer (middle of the figure). In this state, when a user selects the third cluster, articles belonging to c_3^0 are displayed as shown in the left-hand side of the figure in the bottom: when a user selects the second cluster, articles belonging to c_2^0 are displayed as shown in the right-hand side of the figure in the bottom.

In the second and third layer, membership values are visualized with small squares in a cell: the number of squares is proportional to its membership value. This function is expected to be useful for comparing the characteristics of articles within the same cluster, as well as for locating articles of interest.

The size of a cell is determined on the basis of either of the following conditions.

- Membership mode: the size of a cell depends on the membership values for other clusters than the closest cluster.
- Count mode: equal-sized cell is assigned to all articles.

It is supposed that the Membership mode is suitable for identifying articles around the border between different clusters. On the other hand, the Count mode is supposed to be useful for confirming the cluster size.

The cluster information view shows the distribution of categories in a cluster selected with the button of the top-left of the interface. It visualizes the ratio of categories assigned to articles in the cluster with a pie chart.

Figure 4 shows an example of the article information view. When a user hovers the mouse cursor over a cell in the second and third layers of the Treemap view, information about the corresponding article is displayed in this view. The displayed information includes its article number, category, metadata such as publication date and length, title, and main text. Its membership values to all clusters are also visualized with a pie chart.

The cluster relation view shows the relationship between different clusters with a chord diagram. An arc

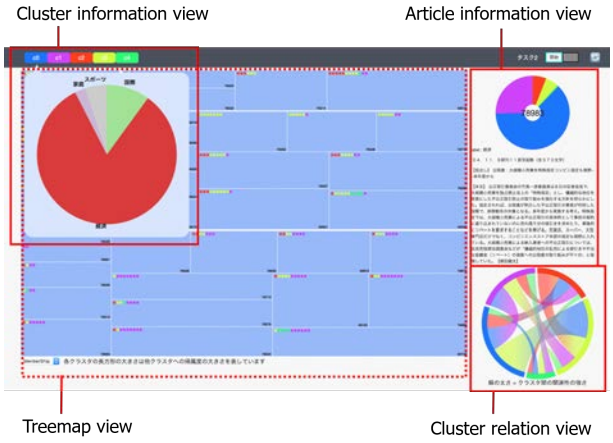

Fig. 2. Screenshot of prototype interface

corresponds to a cluster, and the width of an edge w_{ij} connecting i,j-th clusters is determined on the basis of L2-norm between cluster centroids.

$$w_{ij} = \frac{|v_i - v_j|_2^{-3}}{\sum_{k \neq l} |v_k - v_l|_2^{-3}}, \quad \ldots\ldots\ldots\ldots \quad (5)$$

where v_i is a centroid of i-th cluster. The length of the arc of i-th cluster $l_a(i)$ is represented as Eq. 6, where R is a radius of the diagram.

$$l_a(i) = \pi R \times \sum_j w_{ij}. \quad \ldots\ldots\ldots\ldots \quad (6)$$

A cluster has longer arc when it has a strong relation with other clusters.

5. Experiment

5.1. Outline

To examine the effectiveness of the proposed method, we asked 10 test participants, who are graduate/undergraduate students in Engineering, to do the following tasks using the prototype interface.

T1: Find an article containing all of the 5 words specified in a question (2 questions).

T2: Find an article, of which category is different from other articles in a specified cluster (1 question).

T3: Find all clusters that have weak relation with other clusters (1 question).

T4: Find all cluster pairs that have strong relationship with each other (1 question).

Two questions of the same type were asked for T1, and one question was asked for each of the remaining tasks. Words specified in Task T1 are selected from different 2 categories, so that an article containing those words tends to exist around border of those categories. That is, this task aims to examine the effectiveness of the proposed

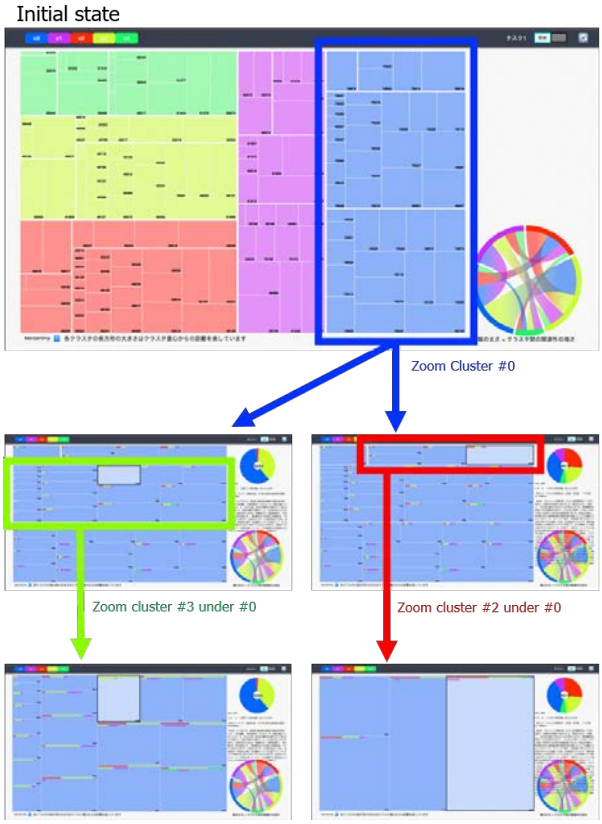

Fig. 3. Example of zooming operation

method for identifying articles that are around the cluster border and difficult to be classified.

A dataset used in the experiment is based on a collection of Japanese news articles published by the Mainichi Newspapers during 2004[4]. 100 articles with 300-800 characters are selected from each of World, Business, Lifestyle&Health, Arts&Entertainment, and Sports: total 500 articles are used in the experiment.

In addition to the prototype interface, the test participants are asked to do the same task as the prototype interface with the following interfaces, which disable some major functions from the prototype interface. The experimental results of those interfaces are compared to evaluate the effectiveness of the major functions.

IF1: Disable the display of membership values with small squares in the cell of the Treemap view.

IF2: Disable the cluster relation view.

The order of using the interfaces as well as the combination of the interface and a dataset is adjusted so as to remove the order effect of using interfaces and the effect of datasets.

After completing all tasks, the test participants were asked to answer a questionnaire, which contains the following questions with 5-point scale (1:bad-5:good).

4. Mainichi Shinbun Kiji Data-syu 2004 Ban: http://www.nichigai.co.jp/index.html

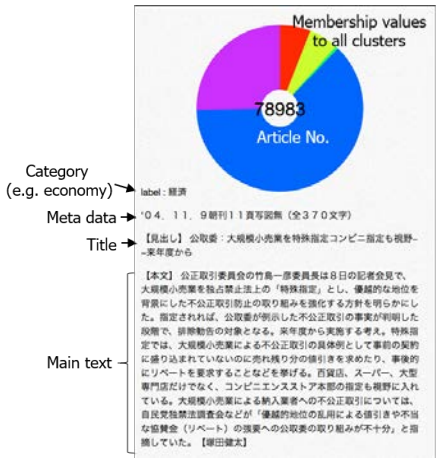

Fig. 4. Article information view

Q1: Quality of clustering result

Q2: Usefulness of displaying membership values with small squares in the cell of the Treemap view

Q3: Usefulness of the cluster relation view

Q4: Visibility of the proposed interface

Q5: Usability of the proposed interface

5.2. Results

Table 1. Comparison of time and interaction: T1

Interface	Average time [s]	Average # of interaction
Proposed	412.70	541.6
IF1	532.28	680.7
IF2	404.79	541.1

Table 2. Comparison of time and interaction: T2

Interface	Average time [s]	Average # of interaction
Proposed	71.46	91.8
IF1	121.45	183.7
IF2	129.50	211

Table 1 and 2 respectively compare average time and the number of interactions spent on completing the tasks T1 and T2. A significant difference was not confirmed among different interfaces. Regarding the task T1, 18 among 20 questions were correctly answered. No difference was observed among different interfaces in terms of accuracy. On the other hand, it was observed that the difficulty of the task depended on the categories of target articles.

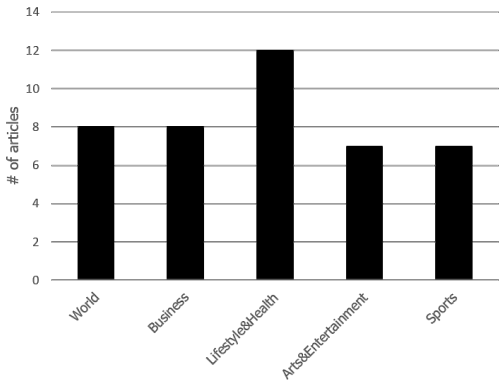

Fig. 5. Categories of articles difficult to find

Figure 5 shows the number of target articles in T1 for which the test participants spent more time and interaction than average for each category. Note that as an article relating with two categories is selected as target, it is counted for both of those categories in the figure. As shown in Fig. 5, the participants tended to spend more time and interaction than average when the target article belongs to Lifestyle&Health. As this category covers broader topics than other categories, it seems to have overlap with other categories.

For the tasks T1 and T2, we asked the test participant to evaluate the validity of the article they found being contained in the cluster with 5-point scale. The correlation between their evaluation and the time spent on completing the task was -0.26317, and the correlation between their evaluation and the number of interactions was -0.25663. A weak negative correlation indicates that it was difficult to find an article that is difficult to classify with Fuzzy c-Means.

Table 3 and 4 respectively compare average time and the number of interactions spent on completing the task T3 and T4. A significant difference was not confirmed among different interfaces.

Regarding the usage of the major functions of the prototype interface when doing tasks T3 and T4, some participants completed the task without using the Treemap view. However, the difference was observed in terms of accuracy of T4 between different interfaces, which is shown in Table 5. The result of ANOVA shows a significant difference among interfaces (p=0.0047). This result indicates the importance of combining the Treemap view and the cluster relation view for accurately understanding the relationship between clusters.

Table 3. Comparison of time and interaction: T3

Interface	Average time [s]	Average # of interaction
Proposed	36.18	58.5
IF1	53.37	74.5
IF2	77.55	190.3

Table 4. Comparison of time and interaction: T4

Interface	Average time [s]	Average # of interaction
Proposed	76.42	97.4
IF1	97.53	228
IF2	107.49	201.6

Table 5. Comparison of Accuracy: T3, T4

Task	Proposed	IF1	IF2
3	0.867	0.85	0.8
4	0.9	0.583	0.367

Figs. 6–10 show the distribution of answers to the questions in the questionnaire. Note that we also collected free comments in addition to 5-scale evaluations.

As shown in Fig. 6, the quality of Fuzzy c-Means was positively evaluated by 8 among 10 test participants.

Fig. 10 shows the usability of the prototype interface is positively evaluated by the majority of the test participant. In particular, positive comments were obtained for the Treemap view, such as it is useful for locating articles belonging to multiple categories, and it is useful when grasping the overview of the dataset by comparing the characteristics of different articles. It corresponds to the result that most of the test participant gave positive evaluations to Q2 (Fig. 7).

Regarding the cluster relation view, they pointed out its effectiveness for understanding the relationship among all clusters all at once. However, it was also pointed out that drawing arcs on the basis of the relative strength as defined with Eq. 5 was inconvenient. It corresponds to the fact that more test participants gave negative evaluations to Q3 and Q4 than Q2. This result suggests the accuracy such as shown in Table 5 could be increased by improving the cluster relation view in the future.

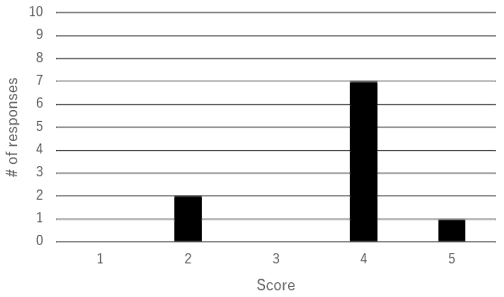

Fig. 6. Response to Q1

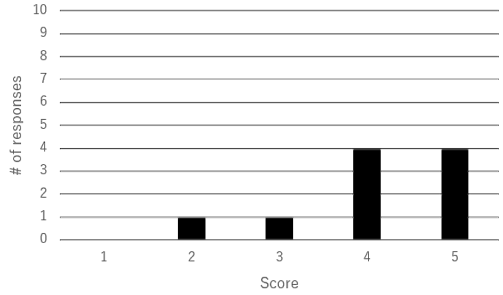

Fig. 7. Response to Q2

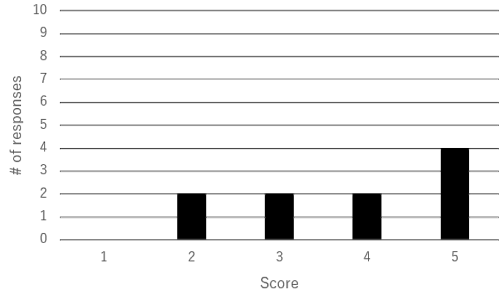

Fig. 8. Response to Q3

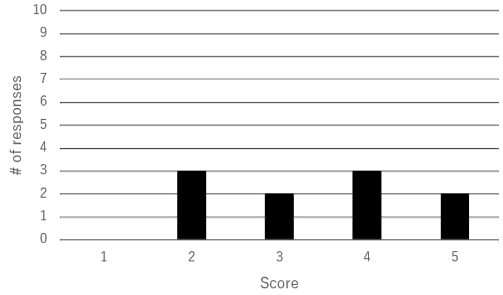

Fig. 9. Response to Q4

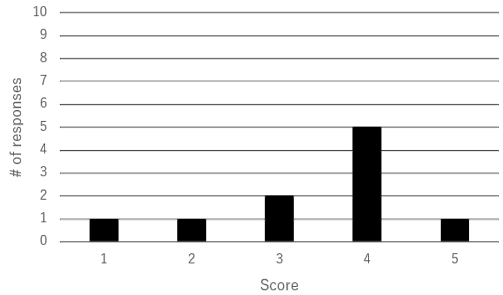

Fig. 10. Response to Q5

6. Conclusion

This paper proposed Treemap-based visualization for supporting cluster analysis of multi-dimensional data. The proposed method visualizes the result of applying Fuzzy c-Means to the target dataset with the Treemap view and the cluster relation view. The result of experiments with test participants showed the effectiveness of the proposed method.

Future works include the improvement of the cluster relation view according to the feedback obtained from the test participants. As noted above, the proposed method could be essentially applied to other kinds of multi-dimensional data than text data: applying the proposed method to different target data is also one of our future works.

References:

[1] S. Akimoto, T. Takahashi, M. Suzuki, Y. Arai, S. Aoyagi, "Human Detection by Fourier Descriptors and Fuzzy Color Histograms with Fuzzy c-Means Method," Journal of Robotics and Mechatronics, Vol. 28, No. 4, pp. 491–499, 2016.

[2] B.B. Bederson, B. Shneiderman, M. Wattenberg, "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies," ACM Transaction on Graphics, Vol. 21, Issue 4, pp. 833–854, 2002.

[3] M. Balzer, O. Deussen, "Voronoi Treemaps," 2005 IEEE Symposium on Information Visualization, pp. 7–14, 2005.

[4] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum Press, New York, 1981.

[5] M. Bruls, K. Huizing, J. van Wijk, "Squarified Treemap," 2005 IEEE Symposium on Visualization, pp. 33–42, 1999.

[6] B. Johnson, B. Shneiderman, "Tree-maps: a Space-filling Approach to the Visualization of Hierarchical Information Structures," 2nd International IEEE Visualization Conference, pp. 284–291, 1991.

[7] K. Matsui, Y. Kageyama, H. Yokoyama, "Analysis of Water Quality Conditions of Lake Hachiroko Using Fuzzy C-Means," JACIII, Vol. 23, No. 3, pp. 456–464, 2019.

[8] Q. Le, T. Mikolov, "Distributed Representations of Sentences and Documents," Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196, 2014.

[9] D. Sacha, L. Zhang, M. Sedlmair, J.A. Lee, J. Peltonen, D. Weiskopf, S.C. North, D.A. Keim, "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis," IEEE Trans. Visualization and Computer Graphics, Vol. 23, No. 1, pp. 241–250, 2017.

[10] E. Sherkat, S. Nourashrafeddin, E.E. Milios, R. Minghim, "Interactive Document Clustering Revisited: A Visual Analytics Approach," IUI2018, pp. 281–292, 2018.

[11] D. Zhang, S. Chen, "Clustering Incomplete Data Using Kernel-based Fuzzy C-means Algorithm," Neural Processing Letters, Vol. 18, pp. 155–162, 2003.

[12] M. Zolkepli, F. Dong, K. Hirota, "Visualizing Fuzzy Relationship in Bibliographic Big Data Using Hybrid Approach Combining Fuzzy c-Means and Newman-Girvan Algorithm," JACIII, Vol. 18, No. 6, pp. 896–907, 2014.