PEOPLE: International Journal of Social Sciences
ISSN 2454-5899

Takamatsu et al., 2018

Volume 4 Issue 1, pp.619-630

Date of Publication: 22nd May, 2018

DOI-https://dx.doi.org/10.20319/pijss.2018.41.619630

This paper can be cited as: Takamatsu, K., Kozaki, Y., Muarakami, K., Bannaka, K., Mitsunari, K., & Nakata, Y. (2018). New Analysis Of Visualization In Eduinformatics Using A Network With Parametric And Nonparametric Correlation Coefficients With Threshold. PEOPLE: International Journal of Social Sciences, 4(1), 619-630.

NEW ANALYSIS OF VISUALIZATION IN EDUINFORMATICS USING A NETWORK WITH PARAMETRIC AND NONPARAMETRIC CORRELATION COEFFICIENTS WITH THRESHOLD

Kunihiko Takamatsu
Faculty of Education, Kobe Tokiwa University, Kobe, Japan
Life Science Center, Kobe Tokiwa University, Kobe, Japan
Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan
ktakamatu@gmail.com

Yasuhiro Kozaki
Faculty of Education, Osaka Kyoiku University, Osaka, Japan
The Center for Early Childhood Development, Education, and Policy Research, The University of Tokyo, Tokyo, Japan
kozaki@cc.osaka-kyoiku.ac.jp

Katsuhiko Muarakami
Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
murakami.ktk@gmail.com

Kenya Bannaka
Department of Oral Health, Kobe Tokiwa College, Kobe, Japan
k-bannaka@kobe-tokiwa.ac.jp

Kenichiro Mitsunari
Faculty of Education, Kobe Tokiwa University, Kobe, Japan
Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan

Regional Liaison Unit, Center for the Promotion of Interdisciplinary
Education and Research, Kyoto University, Kyoto, Japan
kmitsunari@kobe-tokiwa.ac.jp

Yasuo Nakata
Faculty of Health Sciences, Kobe Tokiwa University, Kobe, Japan
y-nakata@kobe-tokiwa.ac.jp

Abstract

Eduinformatics, a new term coined by us, is a field that combines education and informatics, and novel techniques will need to be developed for this field. Earlier, we developed a new visualization method to visualize the curriculum of Kobe Tokiwa University using multidimensional scaling (MDS) and a scatter plot. In this study, our focus is on methods to analyze the relationships between answers to questions in eduinformatics questionnaires. MDS methods are very useful, but have limitations in that their results are difficult to interpret. To facilitate the interpretation of these results, we develop a new visualization method using a network with both parametric and non-parametric correlation coefficients with a threshold (VNCC). VNCC has nine steps. We apply the VNCC method to research on nursing education, and provide an example of the visualization of the result. VNCC methods will be useful in dealing with qualitative research in eduinformatics.

Keywords

Dimension Reduction, Correlation Coefficient, Visualization, Eduinformatics

1. Introduction

In recent years, in the field of artificial intelligence (AI), machine learning and deep learning have made rapid progress. Furthermore, the new field of “data science” is expanding owing to the increase in the amount of data. Since 2011, in the fields of higher education, new fields called “learning analytics” have emerged.

We believe that the current environment surrounding higher education is similar to an older era of life science. After the emergence of bioinformatics, life science became more evidence-based, emphasizing collaboration between biology and informatics. There are indications that a similar revolution, wherein research and data analysis play a crucial role, is set to take place in the field of education as well. Members of our team specialize in different fields such as higher education, nursing, mathematics, bioinformatics, and social welfare. Therefore, we can collaborate and apply new methods that are suitable for education. Recently, we have also encouraged collaboration between many researchers in Kobe Tokiwa University. Their research interests lie in interdisciplinary fields. Further, increasing the abstraction degree to six groups, we unite our research to come up with a novel concept. We named the concept "eduinformatics," that is, a combination of education and informatics. Eduinformatics connects the artistic and scientific fields (Figure 1). Artistic fields involve education, i.e. teaching and learning, while scientific fields involve institutional research, statistics, machine learning, evidence-based research, and informatics. Eduinformatics applies scientific analysis to education. Based on our research, we believe that eduinformatics for higher education will lead to a higher quality of higher education.

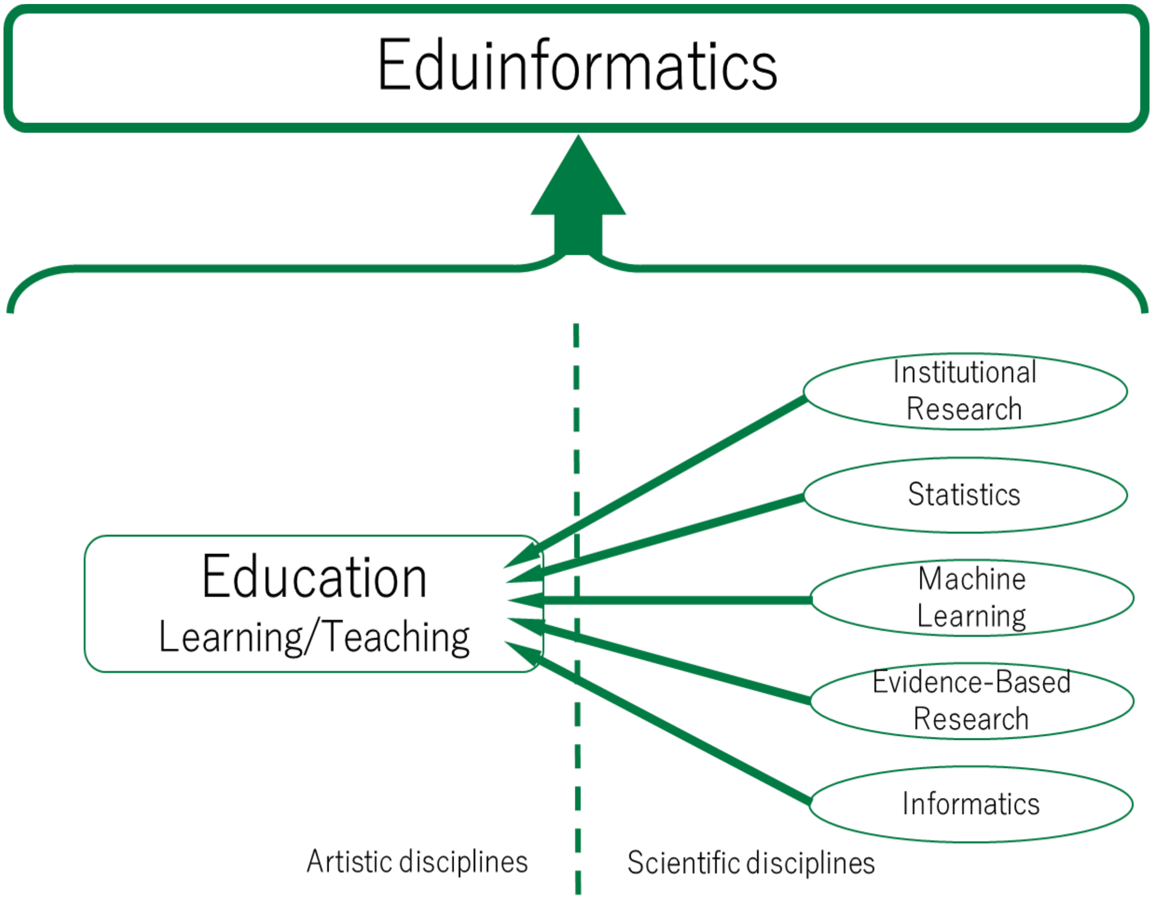

Figure 1: *Concept of Eduinformatics: Eduinformatics combines artistic and scientific disciplines. Scientific disciplines involve institutional research, statistics, machine learning, evidence-based research, and informatics. Artistic disciplines involve teaching and learning. This figure is referred from the article by Takamatsu, Murakami, Kirimura, et al., 2018.*

Methods such as conventional association analysis, crosstab analysis, factor analysis, cluster analysis, logistic regression analysis, linear regression analysis, principal component analysis, and independence tests are used to examine relationships among data. In this study, our focus is on dealing with the relationships between results of questions asked in a questionnaire. Conventionally, correlation coefficients are expressed in the form of a table, or more precisely, a correlation matrix. There are two kinds of correlation coefficients. The first is the well-known Pearson's correlation coefficient (*r*) in parametric analysis. The second is Spearman's rank correlation coefficient (*ρ*) in nonparametric analysis. Even though there is a difference between parametric and nonparametric coefficients, their basic concept is the same.

The definition of the parametric Pearson correlation coefficient is as follows:

$$\begin{aligned} r &= \frac{s_{xy}}{s_x s_y} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} \end{aligned} \tag{1}$$

By extending the concept, the correlation coefficient can be regarded as the angle (cosine) between two vectors in a general vector space. That is, the correlation coefficient coincides with the concept of the cosine function. In fact, this is trivial because the minimum and maximum values of the correlation coefficient coincide with the minimum (−1) and maximum (1) values of the cosine function, respectively. The cosine function (correlation coefficient) is easier to calculate than the Kullback–Leibler divergence, which was developed in 1951. It is used to strictly compare two distributions (Kullback & Leibler, 1951).

When P and Q are discrete probability distributions, the Kullback–Leibler divergence is defined as follows:

$$D_{KL}(P \parallel Q) = \textstyle\sum_i P(i) \log \dfrac{P(i)}{Q(i)} \tag{2}$$

Moreover, when P and Q are continuous probability distributions, the divergence is defined as follows:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{P(i)}{Q(i)} dx \tag{3}$$

The correlation coefficient, the angle (cosine) between two vectors in a vector space, and the Kullback–Leibler divergence are unlikely to satisfy the distance axioms in a general mathematical sense (non-negative, non-degradability, symmetry, and triangular inequality). When any x, y, z belonging to set X satisfy the following four arbitrary conditions against two real variable functions d: X × X → R defined on set X, d is referred to as distance and (X, d) is referred to as the distance space.

(1) non-negative: $d(x, y) \geq 0$

(2) non-degradability: $x = y \Rightarrow d(x, y) = 0$

(3) symmetry: $d(x, y) = d(y, x)$,

(4) triangular inequality: $d(x, y) + d(y, z) \geq d(x, z)$

The correlation coefficient and cosine function can be negative because their minimum value is −1, and the Kullback–Leibler divergence is not symmetric (exchange rule or exchangeable). Thus, the correlation coefficient and Kullback–Leibler divergence are not distances.

Therefore, when practically performing the visualization method (described later), it is necessary to map the correlation coefficient and Kullback–Leibler divergence to the distance space.

The non-negativity of the correlation coefficient can be ensured by subtracting it from 1, and it can be mapped to the distance space. In this case, the maximum and minimum values of the correlation coefficient will become 2 and 1, respectively. As non-degradability, symmetry, and triangular inequality are already established, the value obtained by subtracting the correlation coefficient from 1 satisfies the distance axioms.

In addition, the Kullback–Leibler divergence can be mapped to the distance space by converting it to the symmetrical Jensen–Shannon divergence. The Jensen–Shannon divergence is defined using the Kullback–Leibler divergence as follows:

$$\mathrm{JSD}(\mathrm{P} \parallel \mathrm{Q}) = \frac{1}{2}\mathrm{D}\Big(\mathrm{P} \parallel \frac{1}{2}(\mathrm{P}+\mathrm{Q})\Big) + \frac{1}{2}\mathrm{D}\Big(\mathrm{Q} \parallel \frac{1}{2}(\mathrm{P}+\mathrm{Q})\Big) \tag{4}$$

As can be seen from the definition, the Jensen–Shannon divergence clearly ensures symmetry; thus, it can be mapped to the distance space.

Conventionally, a correlation matrix has been visualized by mapping the correlation coefficient to a distance space and then reducing the dimension of the distance to a low dimension (i.e. two dimensions). The reason for mapping to the distance space is that it is not possible to mathematically reduce the dimension unless it is in the distance space. Reducing a dimension broadly implies that a point group distributed in an ultra-high-dimensional space is mapped to a low-dimensional space while maintaining the distance relation with each element at high probability.

The methods of reducing dimensions can generally be classified into two types, i.e., linear and nonlinear. Linear dimensionality reduction methods have been used for a long time. They include random projection (Bingham, Bingham, Mannila, & Mannila, 2001), principal component analysis (Pearson, 1901), and linear discriminant analysis (Fisher, 1936). However, nonlinear dimensionality reduction methods have only been developed since 2000. A few examples of these methods are Isomap (Tenenbaum, De Silva, & Langford, 2000), locally linear embedding (Roweis & Saul, 2000), modified locally linear embedding (Zhang & Wang, 2006), Hessian eigenmapping (Donoho & Grimes, 2003), local tangent space alignment (Zhang & Zha, 2004), multidimensional scaling (MDS) (Kruskal, 1964), and t-distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten & Hinton, 2008). Recently t-SNE method used in biology (Karaiskos et al., 2017).

In most cases, the (x, y) component obtained through dimensionality reduction is visualized using a scatter plot, which was developed by John Herschel in the 18th century. In this case, it is difficult to understand the meaning of the x and y axes. For simplicity, the x and y axes are images corresponding to the first and second axes of principal component analysis, respectively. To be precise, the x and y axes are derived through calculation when dimensionality is reduced mathematically. Therefore, it is considerably difficult for a person who does not understand the details of dimensionality reduction to accurately understand a visualized figure.

Furthermore, when dimensions are reduced, the distance in the high-dimensional space is reduced; thus, we must consider the loss of information volume. In addition, it should be noted that in the case of nonlinear dimensionally compressed t-SNE, it is unknown how the distance between two elements in two dimensions should be interpreted. If distance between two elements is low in two dimensions, it seems likely that there is a relationship between the two elements; however, this is mathematically incorrect. Furthermore, the closeness between grouped members can be understood from a figure visualized in two dimensions; however, it is difficult to comprehend the closeness if the elements are closer to each other in the original higher dimension as compared to in the figure visualized in two dimensions.

Clustering must be performed to find the distance between groups. Clustering involves decomposing a set of classification objects into subsets based on a certain rule. In general, clustering can be broadly divided into hierarchical and nonhierarchical methods.

Hierarchical methods include the nearest neighbor method (single linkage method), furthest neighbor method (complete linkage method), group average method, and Ward's method. Nonhierarchical methods include the k-means method (k average method).

It is necessary to visualize the results of clustering. For example, hierarchical clustering results are visualized using a phylogenetic tree in biology and a heat map in gene expression analysis. In this case, unlike the visualization of the dimensionality reduction methods described above, it is difficult to interpret results unless the visualization method and clustering method are known.

Therefore, in order to investigate the relationship between data, we have developed a new analysis method, which is referred to as the "visualization using network with parametric and nonparametric correlation coefficients (VNCC)" method.

As a method for visualizing a group of correlation coefficients, the VNCC method performs visualization using a network without a scatter plot or conventional dimensionality reduction and clustering.

2. VNCC Method, an Example, and Discussion

2.1 Research Objective

In the VNCC method, visualization is performed not through dimensionality reduction and clustering, but using the correlation matrix itself. Only correlation coefficients in a correlation matrix with a p value lower than the significance level are strongly correlated. The

VNCC method is a new analytical method that classifies elements as a weak correlation and visualizes it using a network.

2.2 VNCC Methods

The theory of networks is mathematically referred to as graph theory, and Euler's (Leonhard Euler) solution of the "problem of Konigsberg" in the 18th century was regarded as one of its origins. In addition, it is known that graph theory was developed in applied mathematics in the mid-20th century. The "small-world" network proposed by Watts and Strogatz is the first application of graph theory to complex networks (Watts & Strogatz, 1998). Since then, it has been applied in various fields including complex networks.

In particular, in the field of life science, Cytoscape (Shannon, Markiel, Ozier, et al., 2003), which is a network visualization tool developed for the analysis of PPI, is being used in fields other than science.

We have developed a new method to visualize the curriculum of Kobe Tokiwa university (Takamatsu, Murakami, Lim, et al., 2017) (Takamatsu, Murakami, Kirimura, et al., 2017).

First, cosine similarity was calculated based on the text information of the syllabus, the dimension of distance between each text was reduced using the MDS method, and a scatter plot was combined and visualized (Takamatsu, Murakami, Lim, et al., 2017). However, the visualization differed from the actual curriculum map in a few cases. Hence, we developed a new method of visualizing the curriculum, which combines the cosine similarity between subjects based on the distribution of "Tokiwa competencies" described in the syllabus from this year. Then, the dimension was reduced using the MDS method, and a scatter plot was created (Takamatsu, Murakami, Kirimura, et al., 2017).

In this study, first, the cosine similarity between subjects was calculated based on the distribution of "Tokiwa competencies." The cosine similarity is defined as follows:

$$\cos(\overrightarrow{a}, \vec{b}) := \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \frac{\sum_{i=1}^{|V|} a_i b_i}{\sqrt{\sum_{i=1}^{|V|} a_i^2} \cdot \sqrt{\sum_{i=1}^{|V|} b_i^2}} \tag{5}$$

Next, the obtained cosine similarity was subtracted from 1, and a forty-dimensional distance matrix mapped to the distance space was calculated. Then, we reduced the dimension to 2 using the MDS method. Next, visualization was carried out using a scatter plot.

We attempted to visualize the network presented in Figure 3 in the paper by (Takamatsu, Murakami, Kirimura, et al., 2017). However, as all elements (members) appear in the figure, the network becomes extremely complicated, and the visualization is not good compared to the visualization performed using the MDS method.

In this study, we could reduce the original members by the p value first, and it was possible to establish a certain threshold with the following three criteria: strong correlation, correlated, and weak correlation. Therefore, unlike the abovementioned paper, the original number was restricted and the dimension was reduced by matching the strength of the correlation with the thickness of the edge (line). Dimensionality reduction was possible using the p value and threshold, and thus, an appropriate visualization could be obtained.

Furthermore, the primary advantage of the VNCC method is that we can easily understand results. As mentioned above, it is difficult to understand visualization performed through dimensionality reduction or clustering. With the development of the VNCC method, the results obtained in this research as an example have made it possible to intuitively understand the connection (i.e., correlation) between all items.

We have earlier performed research on nursing education. We now present the result of this study, using the VNCC method. In the study, we distributed a questionnaire comprising seventeen items amongst students who had graduated from a nursing course. In this case, we obtained ${}_{17}C_2$=136 correlation coefficients between all items. Whole items are on the nominal scale or ordinal scale. Using the VNCC method, we obtain the result shown in Figure 2. In this figure, we have reduced 136 correlation coefficients to 32 correlation coefficients and 17 items to 16 items by threshold of p value. It can be seen that the results of the VNCC method are intuitively more comprehensible than those of the MDS or clustering methods.

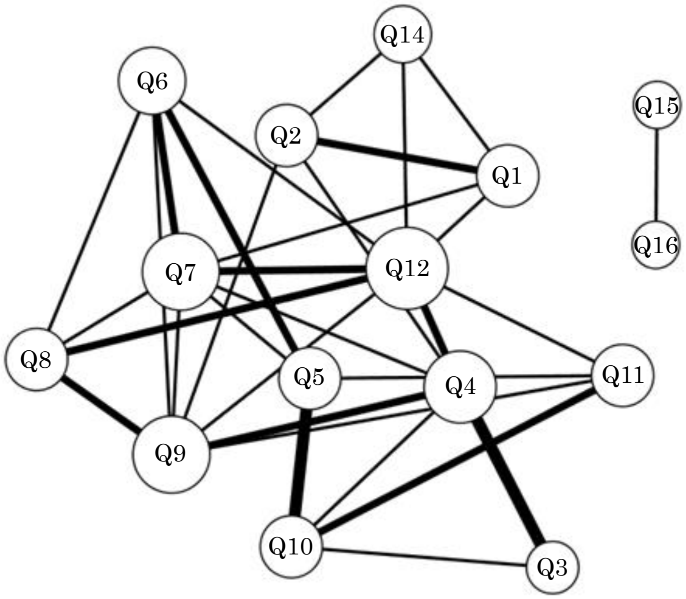

Figure 2: *Example of result using VNCC method.*
This figure is referred from the article by Shoji, Ozaki, Sasao, et al., 2018.

3. Conclusion

In this study, we use both parametric and nonparametric correlation coefficients simultaneously to draw a figure using network analysis. However, it was unclear whether the scale from −1 to 1 is the same between parametric and nonparametric correlation coefficients. In future, we will investigate the relationship between parametric and nonparametric correlation coefficients. When the scale from −1 to 1 is different between parametric and nonparametric correlation coefficients, it would be better to use only nonparametric correlation coefficients, instead of using both parametric and nonparametric correlation coefficients. VNCC methods will be useful in dealing with qualitative research in eduinformatics (Shoji, Ozaki, Sasao, et al., 2018; Hama, Takamatsu, Nakata, & Adachi, 2018).

References

Bingham, E., Bingham, E., Mannila, H., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. *International Conference on Knowledge Discovery and Data Mining (KDD)*, 245-250. http://doi.org/10.1145/502512.502546

Donoho, D. L., & Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, *100*(10), 5591-5596. http://doi.org/10.1073/pnas.1031596100

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. http://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Hama, S., Takamatsu, K., Nakata, Y., & Adachi, R. (2018). Relationship between lip closing force and oral function in healthy women college students. *The Journal of Educational Conference on All Japan Colleges of Dental Hygiene*, *7*, 21-28.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1-27. http://doi.org/10.1007/BF02289565

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., … Zinzen, R. P. (2017). The Drosophila embryo at single-cell transcriptome resolution. Science, 358(6360), 194–199. http://doi.org/10.1126/science.aan3235

Kullback, S., & Leibler, R. A. (1951). JSTOR: The Annals of Mathematical Statistics, Vol. 22, No. 1 (Mar., 1951), pp. 79-86. *The Annals of Mathematical Statistics*. Retrieved from http://www.jstor.org/stable/10.2307/2236703%5Cnpapers2://publication/uuid/A65B3271-44DC-42DC-87B5-A0C3A91EBFA8

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space*. *Philosophical Magazine Series 6*, *2*(11), 559-572. http://doi.org/10.1080/14786440109462720

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323-2326. http://doi.org/10.1126/science.290.5500.2323

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498-2504. http://doi.org/10.1101/gr.1239303

Shoji, Y., Ozaki, Y., Sasao, H., Takamatsu, K., & Nakata, Y. (2018). Significance of pediatric nursing practice in which graduates of our university as clinical nurses teach: From the results of questionnaire survey for students. *Bulletin of Kobe Tokiwa University*, *11*, in press.

Takamatsu, K., Murakami, K., Kirimura, T., Bannaka, K., Noda, I., Yamasaki, M., Lim, R-J. W., Mitsunari, K., Tadashi, N., & Nakata, Y. (2017). A new way of visualizing curricula using competencies: Cosine similarity, multidimensional scaling methods, and scatter plotting. *Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress On. IEEE*, http://doi.ieeecomputersociety.org/10.1109/IIAI-AAI.2017.29.

Takamatsu, K., Murakami, K., Lim, R.-J. W., & Nakata, Y. (2017). Novel visualization for curriculum in silico using syllabus by a combination of cosine similarity, multidimensional scaling methods, and scatter plot: Dynamic curriculum mapping (DCM) for syllabus. *Bulletin of Kobe Tokiwa University*, *10*, 99-106, http://doi.org/10.20608/00000396.

Takamatsu, K., Murakami, K., Kirimura, T., Bannaka, K., Noda, I., Lim, R-J. W., Mitsunari, K., Seki, M., Matsumoto, E., Bohgaki, M., Imanishi, A., Omori, M., Adachi, R., Yamasaki, M., Sakamoto, H., Takao, K., Asahi, J., Nakamura, T., & Nakata, Y. (2018). "Eduinformatics": A new education field promotion. *Bulletin of Kobe Tokiwa University*, *11*, 27-44.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319-2323. http://doi.org/10.1126/science.290.5500.2319

Van Der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605. http://doi.org/10.1007/s10479-011-0841-3

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. http://doi.org/10.1038/30918

Zhang, Z., & Wang, J. (2006). MLLE: Modified locally linear embedding using multiple weights. *Advances in Neural Information Processing Systems*, 1593–1600. Retrieved from http://cognet.mit.edu/library/books/mitpress/0262195682/cache/chap200.pdf

Zhang, Z. Y., & Zha, H. Y. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University*, *8*(4), 406–424. http://doi.org/10.1007/s11741-004-0051-1