DOI: 10.33965/celda2019_201911L045
16th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2019)

CLUSTERING OF LEARNERS BASED ON KNOWLEDGE MAPS

Akira Onoue, Atsushi Shimada, Tsubasa Minematsu and Rin-Ichiro Taniguchi
Kyushu University, Japan

ABSTRACT

This study aimed to cluster learners based on the structures of the knowledge maps they created. Learners drew their own knowledge maps to reflect their learning activities. Our system collected individual knowledge maps from many learners and clustered them to generate an integrated version of the knowledge maps of each cluster. We applied the graph analysis method to extract important keywords from the knowledge map. The results of the analysis showed that the utilization of the knowledge map helped to improve lectures and grasp the learners' level of understanding. We conducted surveys asking course managers to evaluate the effectiveness of the integrated knowledge maps of learners included in the cluster and received both positive and negative responses.

KEYWORDS

Knowledge Map, Similarity, Pagerank, Netsimile, Clustering, Infinite Relational Model

1. INTRODUCTION

Research on the cognitive perspective of learning has been conducted for a long time in the educational psychology field. Cognitive learning in a learning process is one of the most important perspectives for successful learning performances. One of the effective cognitive learning support tools is a concept map, which allows learners to make a map connecting important concepts and their ideas. It enhances the use of cognitive learning strategies (e.g., Fiorella and Mayer, 2017) and productive discussions in collaborative learning settings (Yamada et al., 2016).

In the highly advanced information technology era, a concept map can play an additional significant role in learning. Log data from a digital concept map tool allows researchers and teachers to track the learning process using log data visualizations such as the concept map construction process (Hsiao and Brusilovsky, 2012). As a similar cognitive tool, a knowledge map can be an effective cognitive learning tool. Knowledge maps focus on the construction of knowledge relationships, such as the linkages among information (Crampes et al., 2006; Balaid et al., 2016); this suggests that a knowledge map is effective at granting access to knowledge in a timely manner, identifying knowledge flow, allowing organizational restructuring, and so on. However, categorization and visualization, such as clustering, are required in order to confirm the knowledge construction process and elucidate the knowledge gap between high and low performers. For example, clustering of both learning process data and learning performance data allows teachers to comprehend effective learning styles (Yin et al., 2019). Clustering data seems to help teachers consider instructional designs and support learners.

This research focuses on knowledge map construction and clustering and aims to develop a teaching support system using knowledge map clustering data.

2. RECONSTRUCTION OF A KNOWLEDGE MAP

A knowledge map is a network in which learned keywords are arranged as nodes, and the relationships between the nodes are indicated using arrows. This section discusses how to reconstruct the individual knowledge maps created by many learners.

First, we will introduce the e-book system and BR-Map tool (Yamada et al., 2018), followed by the reconstruction and analytical strategy of knowledge maps. Each learner creates his/her own knowledge map using the BR-Map tool. The words in the nodes of the BR-Map are automatically extracted from the e-book system by referring to the areas highlighted by the learner. The highlighted words or sentences become candidates for nodes in the BR-Map system. Finally, the learners create their own knowledge map by arranging nodes and drawing link nodes.

Second, we executed node and link processing. By using the BR-Map tool, we acquired the knowledge maps created by learners. These knowledge maps contain some nodes corresponding to a single word, sentence, number, symbol, and so on. Ideally, each node should have one word (keyword) to represent a knowledge point. Therefore, we adopted a text mining process to identify the same keywords in sentences proposed by Onoue et al. (2019). There were two steps in determining the nodes in the integrated knowledge map:

1. Morphological analysis by MeCab (Kudo, 2006) with mecab-ipadic-NEologd (Sato et al., 2016).
2. Integration of nodes with similar words based on the normalized Levenshtein distance.

As a result of node processing, each node had one word. In the next step, we analyzed the links between nodes, and separated or integrated them according to the condition of the connected nodes. After node and link processing, we acquired the individual knowledge map of each learner (refer to Onoue et al. [2019] for details about node and link processing).

Third, we made clusters of the learners based on their individual knowledge maps. The clustering algorithm is introduced in detail in section 3.

Lastly, we integrated the knowledge maps of learners included in each cluster. We used a method to create the integrated knowledge maps based on Onoue et al. (2019). The integrated knowledge maps show how course contents are organized and remembered by learners. This information is important for both teachers and learners to reflect their teaching and learning activities, respectively. On the integrated knowledge map, a link between nodes is represented as a weighted link. If two or more learners establish the same link between a certain combination of words, the weight of the link is increased based on the number of learners. The centrality algorithm provides helpful information for understanding the relationships between words that have strong connections with each other.

3. CLUSTERING OF KNOWLEDGE MAPS

A knowledge map is a network portraying how a learner organizes and understands what they have learned. The purpose of this study is to cluster learners based on knowledge maps. Clustering results are useful for helping teachers to understand the patterns of learners' understanding. Providing appropriate information for each learner based on a clustering result can improve learner understanding.

3.1 Similarity between Learners' Individual Knowledge Maps

We calculated the similarity between the individual knowledge maps created by learners using a method based on NetSimile (Berlingerio et al., 2013). The knowledge maps were directed graphs with labeled nodes. Therefore, we needed to take into account the direction of the links and the meaning of the nodes when the similarity between learners' individual knowledge maps was calculated; however, NetSimile was developed for undirected graphs with unlabeled nodes. Therefore, we extended NetSimile so that it could be used to make comparisons of the knowledge maps.

3.1.1 NetSimile

NetSimile is a method for calculating the similarity between two graphs. The similarity of these graphs is defined as the similarity of their "signature" feature vectors. NetSimile has three steps: feature extraction, feature aggregation, and comparison.

During feature extraction, we generated a set of structural features for each node based on its local and egonet features. An egonet is a subgraph consisting of a focus node, and the nodes have a link with a focus node. The original NetSimile algorithm had the following seven features:

$d_i = |N(i)|$: degree of node i, $N(i)$ denotes the neighbors of node i.

c_i: clustering coefficient of node i defined as the number of triangles connected to node i over the number of connected triples centered on node i.
$\overline{d}_{N(i)} = \frac{1}{d_i}\sum_{\forall j \in N(i)} d_j$: the average number of node i's two-hops-away neighbors.
$\overline{c}_{N(i)} = \frac{1}{d_i}\sum_{\forall j \in N(i)} c_j$: the average clustering coefficient of $N(i)$
$|E_{ego(i)}|$: the number of edges in node i's egonet $ego(i)$.
$|E^o_{ego(i)}|$: the number of links between $ego(i)$'s nodes and the outside nodes of $ego(i)$.
$|N(ego(i))|$: the number of neighbors of $ego(i)$.

During feature aggregation, NetSimile generates a $nodes \times features$ matrix F_{g_j}, for each graph $g_j \in G = \{g_1, g_2, \ldots, g_k\}$. Then, NetSimile calculates the following five values in each feature (i.e., each column of F_{g_j}) to produce "signature" feature vectors $\boldsymbol{s}_{g_j}$: median, mean, standard deviation, skewness, and kurtosis. Therefore, each graph g_j is represented by five parameters.

Lastly, during the comparison step, we calculated the Canberra distance $d_{Can}(P,Q) = \sum_{i=1}^{d} \frac{|P_i - Q_i|}{P_i + Q_i}$ between the feature vectors P and Q.

3.1.2 NetSimile for Directed Graphs

To cluster the individual knowledge maps created by each learner, we calculated the similarity between the directed graphs with labeled nodes. As mentioned above, NetSimile is a method for calculating the similarity between two undirected graphs with unlabeled nodes. However, we needed to distinguish similarly structured graphs if the nodes had different labels. Therefore, we extended NetSimile to handle directed graphs with labeled nodes. The similarity of these graphs was defined based on the similarity of their "signature" feature vectors. The algorithm has three steps: feature extraction, vectorization, and comparison.

During feature extraction, we took the links' direction into account. Since the learners were instructed to draw links from the upper concept to the lower concept on their individual knowledge map, not only the existence of the link but also its direction was important information. We defined the following 10 features:

$d_i^{in} = |N_{in}(i)|$: income degree of node i, $N_{in}(i)$ denotes the neighbors with incoming links to node i.
$d_i^{out} = |N_{out}(i)|$: outgoing degree of node i, $N_{out}(i)$ denotes the neighbors with outgoing links from node i.
c_i: clustering coefficient of node i defined as the number of triangles connected to node i over the number of connected triples centered on node i.
$\overline{d}_{N_{in}(i)} = \frac{1}{d_i}\sum_{\forall j \in N_{in}(i)} d_j$: the average number of node i's two-hops-away neighbors with incoming links to neighbors of node i.
$\overline{d}_{N_{out}(i)} = \frac{1}{d_i}\sum_{\forall j \in N_{out}(i)} d_j$: the average number of node i's two-hops-away neighbors with outgoing links from neighbors of node i.
$\overline{c}_{N(i)} = \frac{1}{d_i}\sum_{\forall j \in N(i)} c_j$: the average clustering coefficient of $N(i)$
$|E_{ego(i)}|$: the number of links in node i's egonet $ego(i)$.
$|E^{in}_{ego(i)}|$: the number of income links to $ego(i)$'s nodes from outside nodes of $ego(i)$.
$|E^{out}_{ego(i)}|$: the number of outgoing links from $ego(i)$'s nodes to outside nodes of $ego(i)$.
$|N(ego(i))|$: the number of neighbors of $ego(i)$.

During the vectorizing step, we generated the "signature" feature vector $\boldsymbol{s}'_{g_j}$. $\boldsymbol{s}'_{g_j}$ by vectorizing (arranging in a row) the feature matrix F_{g_j}. Both the original and our proposed methods needed to match the number of dimensions of the feature vectors for calculating the feature vector distance. In the original NetSimile algorithm, the number of rows of each feature matrix F_{g_j} is different for each graph because each graph has a different number of nodes. In order to make it possible to calculate the distance between the feature vectors, the number of dimensions was aligned in the feature aggregation step. On the other hand, the aggregation step cannot handle word information on the nodes. Therefore, we avoided the aggregation strategy and took the other approach to consider the word information on each node in our proposed method.

ISBN: 978-989-8533-93-7 © 2019

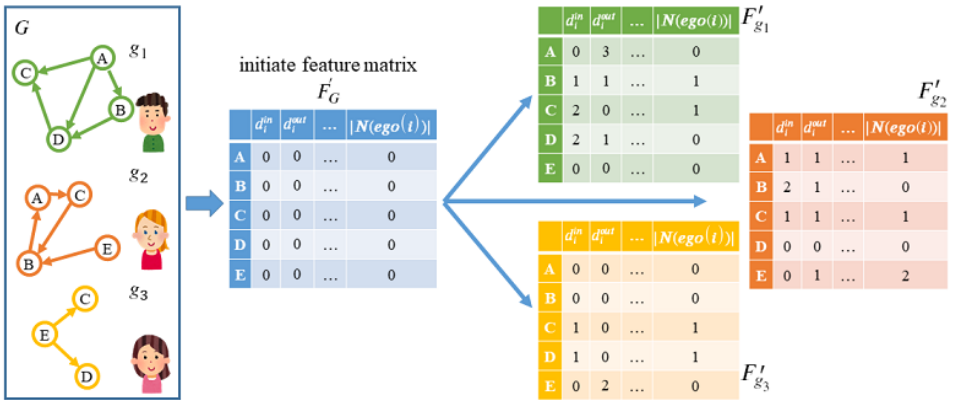


Figure 1. Example of creating a feature matrix

We made a dictionary vector , which contained all of the words in $\boldsymbol{G}$ without duplicates. As a result, the number of rows of feature matrix F'_G became a list of all words, as shown in Figure 1 (A, B, C, D, and E are the lists of words extracted from all graphs without duplication). We gave a score of zero to all features corresponding to a word when a graph did not contain the word in the nodes. For instance, g_1 did not have a node with the word "E," so the scores in the fifth row in F'_{g_1} were all zero. This approach enabled the alignment of the length of rows for all graphs, resulting in generating the same dimensional feature vectors. The feature vector of each graph was represented by the concatenation of row vectors.

Lastly, in the comparison step, we calculated the cosine distance $d_{Cos}(P,Q) = \frac{\sum_{i=1}^{|V|} P_i Q_i}{\sqrt{\sum_{i=1}^{|V|} P_i^2}\sqrt{\sum_{i=1}^{|V|} Q_i^2}}$.

3.2 Clustering Learners using the Infinite Relational Model

After calculating the similarity between each learner's knowledge map, we created a relation matrix $\boldsymbol{R} = (r_{ij})_{i,j=1}^{l}$ so that $r_{ij} = 1$ if the similarity between learner i's knowledge map and learner j's knowledge map is over the threshold (th). In the other case, we set it so that $r_{ij} = 0$. l denoted the number of learners. $\boldsymbol{R}$ is a symmetric matrix indicating whether each learner's knowledge map has strong relationships.

Lastly, the infinite relational model (IRM) (Kemp et al., 2006) was applied to $\boldsymbol{R}$ in order to classify learners based on their knowledge maps. IRM, which is based on a nonparametric Bayesian model, can estimate the number of hidden clusters from binary relational data (refer to Kemp et al. [2006] for the detailed algorithm of the IRM).

4. EXPERIMENT

4.1 Subjects and Course

We conducted the experiments during the university education course. The main themes of the course were basic skills, laws, and ethics related to cybersecurity. The course was conducted over eight weeks from April to June 2018. In total, 98 first-year students created an individual knowledge map for this course. After the final lecture, we asked the learners to create knowledge maps for the purpose of reflecting what they had learned over the eight weeks.

4.2 Evaluation of the Sub-Maps

After calculating the similarity between each learner's knowledge map, we created a relation matrix $\boldsymbol{R}$ shown in Figure 2. In this research, we set the threshold of similarity (th) to be 0.3 when creating a relation matrix. The yellow part in Figure 2 corresponds to the combination of learners who obtained similarities larger than th. The red and blue lines represent the boundaries of clusters in the row and column

directions. The learners were classified into seven clusters in the row direction and six clusters in the column direction. In the rectangular area surrounded by blue and red lines, the larger the proportion of yellow elements, the stronger the connection between the learners corresponding to the rows and columns. Although there were some clusters with strong similarities, such as the right block of cluster "A" in Figure 2, we focused only on the clusters that contained the same learners in both the rows and columns as much as possible. As a result, we selected four clusters (A, B, C, and D in Figure 2 to evaluate the result of the clustering. To analyze the characteristics of each cluster, we made an integrated knowledge map for each cluster.

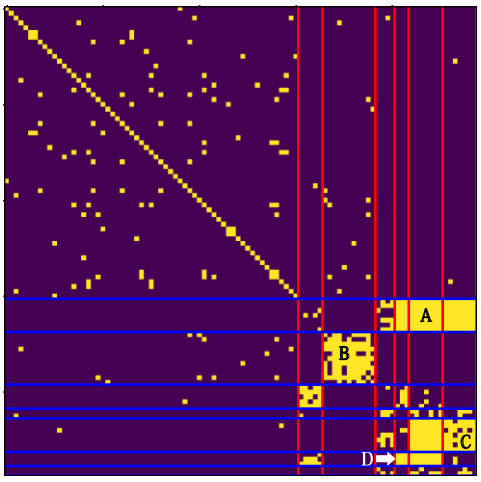


Figure 2. Relation matrix sorted by the IRM

Figure 3. Correspondence between lecture themes and node colors

We administered a questionnaire to the course managers who designed the course and provided the lecture materials, in order to evaluate the effectiveness of sub-map k, which is the integrated knowledge map of learners included in the cluster. A larger-sized node represented an important node, which means that many learners drew links to/from the node. Additionally, the node's color corresponded to the lecture that had learning materials in which the word was frequently used (shown in Figure 3). In order to evaluate the clustering result, we constructed the whole integrated knowledge map shown in Figure 4, and sub-maps k_A~k_D shown in Figure 5. The whole integrated knowledge map is a combination of the individual knowledge maps of all learners. Each sub-map had structural features. k_A mainly consisted of nodes about copyright, and there weren't almost links between nodes of different lectures. k_B consisted of individual knowledge maps of the most learners. There were mainly nodes of introduction and cryptography. k_C mainly consisted of nodes of introduction, ethics and copyright. k_D consisted only of nodes about copyright. There were links between the nodes of introduction and the nodes of other lectures. After presenting the integrated knowledge map and sub-maps to the course manager, we asked them to fill out the questionnaire.

Table 1 shows the questionnaire about the integrated knowledge maps and the responses. Each course manager answered each question according to a five-grade evaluation system. Q1–Q3 asked about the readability of k_A~k_D. Q4–Q6 asked whether k_A~k_D matched the course managers' aim for the course. Q4 and Q5 asked them to consider whether it would help to improve classes in terms of supporting education. Q7–Q9 asked about how to use sub-maps. In particular, we investigated whether the course managers wanted to compare sub-maps or the whole integrated knowledge map with the sub-maps. Q10 and Q11 asked whether utilizing the knowledge map would be useful for improving the content of the lectures, and for grasping the learners' level of understanding. Q12 asked whether the course managers wanted to use sub-maps in their educational activities.

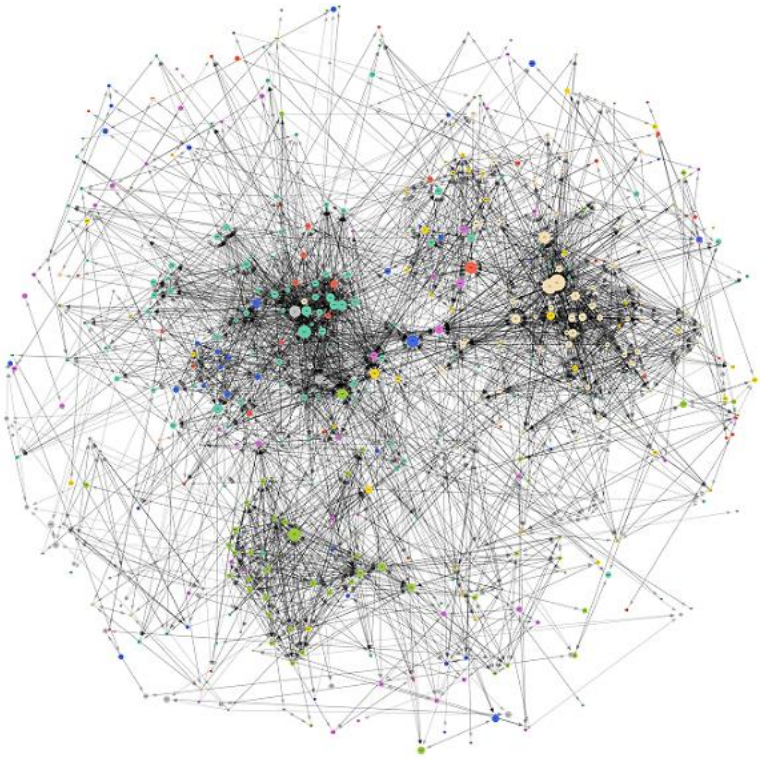
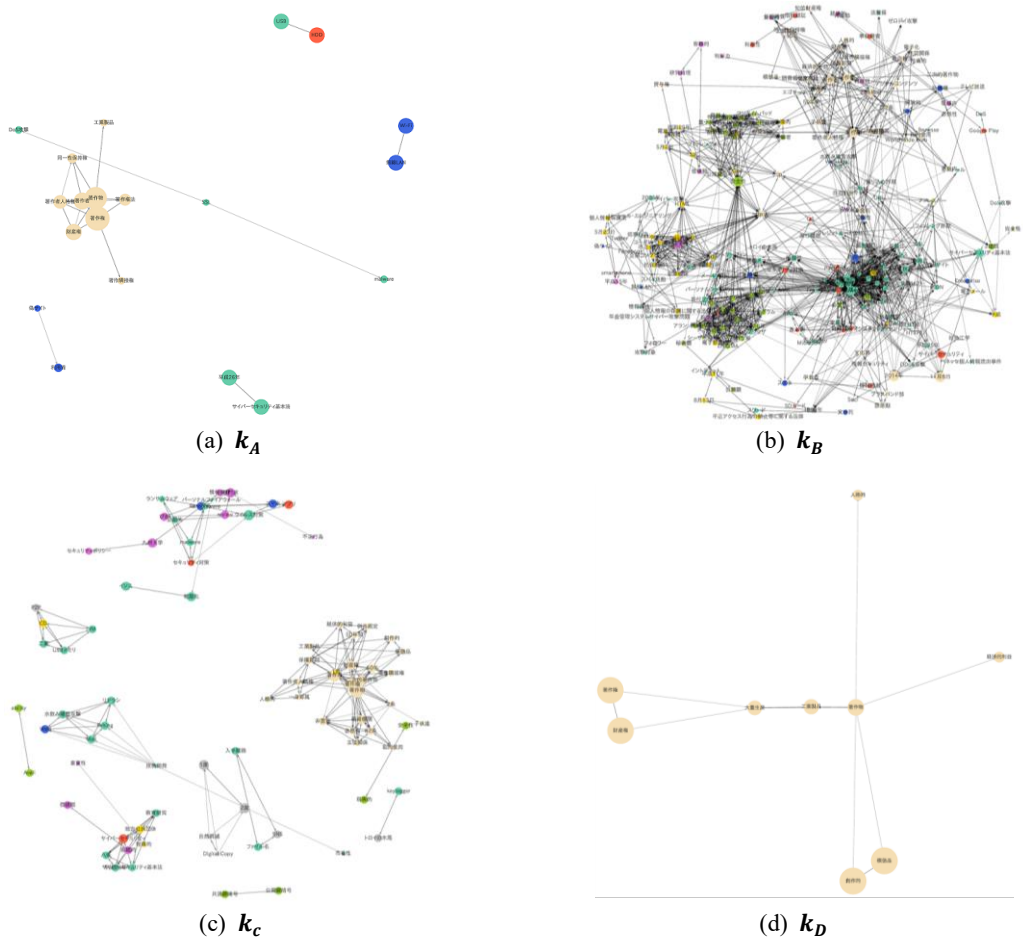


Figure 4. The whole integrated knowledge map

Table 1. Results of the questionnaire about the clustered integrated knowledge map

	Question	Evaluation				
		Strongly agree	Agree a little	Neither agree nor disagree	Disagree a little	Strongly disagree
1	From the sub-maps, it is possible to grasp which part of the lecture contents the learners understand.	0	4	0	0	0
2	It is easier to recognize the understanding of each learner than the whole knowledge map.	2	2	0	0	0
3	I can identify the relationships between the contents of each lecture.	0	2	1	1	0
4	It is supposed to be divided into the clusters.	0	0	3	1	0
5	The result of the sub-map matches the purpose of this class.	0	0	4	0	0
6	The nodes I consider to be important are larger than others.	0	0	1	3	0
7	I want to know the same part between sub-maps.	0	1	1	2	0
8	I want to know the different parts of sub-maps.	0	1	1	2	0
9	I want to know which part the sub-maps constitute in the whole knowledge map.	1	2	0	1	0
10	The sub-map can be used effectively for lecture improvement.	0	3	1	0	0
11	It is useful for grasping the students' understanding.	0	4	0	0	0
12	I want to use the sub-map.	0	3	1	0	0



(a) $\boldsymbol{k_A}$ (b) $\boldsymbol{k_B}$

(c) $\boldsymbol{k_c}$ (d) $\boldsymbol{k_D}$

Figure 5. Sub-maps of clusters A, B, C, and D

We received both positive and negative responses. The question on the importance values of each node and the comparison between sub-maps mainly received negative responses. The answers to Q6 showed that the nodes with high importance were not considered as important by the course managers. We guess that the course managers had difficulty in comparing the importance of the sub-maps because the sizes of the nodes were not regulated among the sub-maps. Besides, the answers to Q7 and Q8 suggested that comparisons among the sub-maps are not necessary for the course managers. In our future work, based on the answers to Q6–Q9, we are going to improve the visualization strategy to emphasize the sub-maps on the whole knowledge map. We expect that this strategy will help course managers to grasp the relationships not only between the whole map and sub-maps, but also among the sub-maps.

On the other hand, the positive results were mainly about the readability of the sub-maps (k_A~k_D). The answers to Q1 and Q3 suggested that the course managers could grasp the students' situations, i.e., which contents the students had interest in from the lectures. Additionally, from the answers to Q2, we found that the sub-maps were more useful than the whole integrated knowledge map used on its own. It turned out that the course managers wanted a summary of the knowledge map rather than all of the information. Moreover, the answers to Q10–Q12 suggested that the course managers thought that utilizing the knowledge map would help to improve the lectures and grasp the learners' level of understanding. Furthermore, the course managers wanted to use sub-maps in their educational activities. We received the following comments regarding improving the sub-maps:

- Sharing the comments of the course managers reading the sub-maps.
- Annotations on the structural features of the sub-maps.

In comparison with the results of the questionnaire mentioned above, the answers to Q4 and Q5 were neither positive nor negative. We will continue with additional interviews to clarify how the course managers interpreted the knowledge maps and the purpose of reading the knowledge maps in our future work.

5. CONCLUSION

This study aimed to cluster learners based on knowledge maps and to develop a teaching support system using knowledge map clustering data. A knowledge map is a network showing how a learner organizes and understands what they have learned. We calculated the similarity between the individual knowledge maps based on NetSimile and classified learners based on their knowledge maps using the IRM. We administered a questionnaire to the course managers who designed the course and provided the lecture materials in order to evaluate the effectiveness of the sub-maps. We received both positive and negative responses. The importance values of each node and comparisons among the sub-maps received negative responses. In contrast, the positive responses were mainly related to the readability of the sub-maps and utilizing the knowledge map.

In future research, we intend to continue developing the analyzation method of the knowledge map. First, it was difficult for the course managers to compare the importance values among the sub-maps because the sizes of the nodes were not regulated among the sub-maps. In addition, the course managers wanted to compare the whole map and sub-maps. Therefore, we are going to improve the visualization system to emphasize the sub-maps on the whole knowledge map. Second, the course managers asked for an explanation of what the sub-maps represented. Therefore, we will consider adding functions such as the automatic generation of annotations to make it easier for users of knowledge maps to understand their structural features. In this study, we manually selected clusters to be visualized as sub-maps. However, when there is a lot of data to be analyzed, manual cluster selection is not desirable. In order to promote the use of knowledge maps for the improvement of teaching methods, we will develop an automatic selection method for the clusters to be visualized.

ACKNOWLEDGEMENT

This work was supported by JST AIP Grant Number JPMJCR19U1 and JSPS KAKENHI Grand Number JP18H04125, Japan.

REFERENCES

Balaid, A. et al., 2016. Knowledge maps: A systematic literature review and direction for future research, *International Journal of Information Management*, 36, pp.451-475

Berlingerio, M. et al., 2013. Network similarity via multiple social theories. *In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagara Falls, Ontario, Canada, pp. 1439-1440.

Crampes, M. et al., 2006. Concept Maps for Designing Adaptive Knowledge Maps, *Information Visualization*, 5(3), pp.211-224.

Fiorella, L., & Mayer, R.E. 2017. Spontaneous spatial strategy use in learning from scientific text, *Contemporary Educational Psychology*, 49, pp.66-79.

Hsiao, I. H., & Brusilovsky, P. 2012. Motivational social visualizations for personalized e-learning. *In European Conference on Technology Enhanced Learning*, Springer, Berlin, Heidelberg, pp. 153-165

Kemp, C. et al., 2006. Learning systems of concepts with an infinite relational model. *In AAAI*, Vol. 3, pp. 5

Kudo, T. 2006. *Mecab: Yet another part-of-speech and morphological analyzer*. [online] Available at: http://mecab. sourceforge.jp. [Accessed 17 Jul. 2019].

Onoue, A. et al., 2019. The Integrated Knowledge Map for Surveying Students' Learning. *In Society for Information Technology & Teacher Education International Conference*, Las Vegas, NV, United States, pp. 838-846.

Page, L., et al., 1999. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Sato, T. et al., 2017. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval. *In Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*. The Association for Natural Language Processing.

Yamada, M. et al., 2016. A computer-supported collaborative learning design for quality interaction. *IEEE Multimedia*, 23(1), pp.48-59.

Yamada, M. et al., 2018. BR-Map: Concept Map System Using E-Book Logs, *Proceedings of 15th International Conference of Cognition and Exploratory Learning in Digital Age*, Budapest, Hungary, pp. 248-254.

Yin, C. et al., 2019. Exploring the Relationships between Reading Behavior Patterns and Learning Outcomes based on Log Data from e-books: A Human Factor Approach, *International Journal of Human-Computer Interaction*, pp.313-322.