International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

An integrated approach for knowledge extraction and analysis in collaborative knowledge construction

Ning Zhang[1] and Fan Ouyang[1*]

*Correspondence: fanouyang@zju.edu.cn

[1] College of Education, Zhejiang University, Hangzhou 310058, Zhejiang, China

Abstract

Collaborative knowledge construction (CKC) involved students' sharing of information, improvement of ideas, and construction of collective knowledge. In this process, knowledge extraction and analysis can provide valuable insights into students' knowledge capacities, depths, and levels in order to improve the CKC quality. However, existing studies tended to extract and analyze knowledge from a single perspective (e.g., the number of certain knowledge types and knowledge structures), which failed to demonstrate the complexity and dynamics of knowledge construction and advancement. To fill this gap, this research designed a series of computer-supported collaborative concept mapping (CSCCM) activities to facilitate students' CKC process and then used an integrated approach (i.e., semantic knowledge analysis combined with learning analytics) to extract, analyze, and understand students' knowledge characteristics and evolutionary trends. Results demonstrated that compared to the low-performing pairs, the high-performing pairs mainly discussed knowledge related to the course content, and their knowledge evolution trend was relatively stable. Based on the results, this research provided analytical implications to extract, analyze, and understand students' knowledge and pedagogical implications to promote students' knowledge construction and advancement.

Keywords: Knowledge extraction and analysis, Knowledge evolution, Semantic knowledge, Integrated approach, Collaborative knowledge construction, Computer-supported collaborative concept mapping

Introduction

The significance of knowledge is emphasized in the current information society (Anderson, 2008; Välimaa & Hoffman, 2008), and relevant practices, e.g., knowledge discovery (Pazzani, 2000), knowledge management (Durst & Zieba, 2018), and knowledge construction (Charlton & Avramides, 2016) have been long studied. Knowledge externalization, as a critical element component in these practices, is regarded as a learner's conscious process of presenting his/her inner knowledge to the public through varied media, e.g., audio, text, images, concept map, etc. (Ifenthaler et al., 2011; Lehmann et al., 2014). Particularly, in collaborative knowledge construction (CKC), students must externalize their knowledge through sharing information and resources, comparing and negotiating disagreements, and synthesizing and co-constructing knowledge (Fischer et al., 2002; Mayordomo & Onrubia, 2015; Zabolotna et al., 2023). Although studies put emphasis on the value of understanding students' knowledge (Ashwin, 2014; Felder & Brent, 2005; Simonsmeier et al., 2022), few studies have addressed the extraction and analysis of student knowledge in the CKC context. Currently, three methods for knowledge extraction are used, namely manual coding, semi-automatic analysis, and automatic analysis. First, existing research has extensively employed manual coding in terms of established knowledge classification frameworks to extract knowledge (Liu et al., 2021; Phillips et al., 2019). However, the traditional manual coding approach often involved labor-intensive and time-consuming work that may produce subjective results. Semi-automatic analysis is the utilization of artificial intelligence (AI) algorithms to train knowledge classification models based on a training corpus, which highly relies on human labelling or intervention. A major disadvantage of semi-automated analysis approach is that the AI trained models may not be suitable for application in other research or educational contexts, which reduces the capacity of generalizability and accuracy of knowledge classification models (Patikorn et al., 2019). Automatic analysis involves the use of advanced artificial intelligence algorithms technologies to automatically learn and infer knowledge types from unlabeled data without human coding and labelling. One of the automatic analysis methods is semantic knowledge analysis, which uses approaches such as semantic network analysis and topic-modeling to extract concepts, ideas, or knowledge based on the linguistic units (e.g., words, sentences, chapters) (Liu & Chen, 2023; Pfiffner, 2021; Wu et al., 2021). Nevertheless, existing analyses tend to extract relevant keywords, define knowledge types in terms of these keywords, and subjectively judge student knowledge capabilities, which require high responsibility, skills, and expertise from the researchers or coders (Jelodar et al., 2019). It is necessary to automatically determine the type of knowledge associated with keywords without manual intervention in order to improve efficiency and accuracy of knowledge extraction and analysis. To fill this gap, this research designed a series of computer-supported collaborative concept mapping (CSCCM) activities to assist students' CKC, and then integrated learning analytics methods with semantic knowledge analysis based on a knowledge base to extract, analyze, and understand students' knowledge construction process. Specifically, this research compared the characteristics and evolutions of semantic knowledge between high-performing pairs and low-performing pairs. Based on the results, this research provided pedagogical implications to promote future instructional practices and analytical implications in the CKC process.

Literature review

Computer-supported collaborative concept mapping as a knowledge externalization means

Grounded upon the socio-cultural perspective of learning (Vygotsky, 1978), collaborative knowledge construction (CKC) emphasizes students' sharing and organization of information, construction, and advancement of knowledge, and establishment of consensus and reflection through peer interactions in groups (Fischer et al., 2002; Mayordomo & Onrubia, 2015; Zabolotna et al., 2023). As a means of CKC, computer-supported collaborative concept mapping (CSCCM) provides students with opportunities to organize and externalize their knowledge, clarify and distinguish concepts, and integrate new knowledge into their prior knowledge (Chiou, 2008; Greene & Azeved, 2010). CSCCM is an effective strategy for externalizing and representing knowledge in the graphical formats, which consists of nodes denoting knowledge concepts and labeled lines representing the relationships between concepts (Novak et al., 1983). Studies have proved that CSCCM has the potential to enhance students' cognitive abilities, develop their higher-order thinking skills, and foster students' deep learning process (Chang et al., 2017; Chu et al., 2019; Sundararajan et al., 2018). The analysis of the knowledge reflected in the concept map enables instructors to gain insight into students' knowledge capabilities and provide guidance for instructional interventions. Through the analysis of knowledge across students with different performances, instructors can identify the knowledge deficiencies of low-performing students, which supports instructors in providing targeted strategies for these students to succeed academically. However, extracting knowledge from concept maps is a challenging and tedious task. There is a lack of fixed standards for the analysis of knowledge from concept maps due to the ill-structured characteristics of CSCCM (Jonassen, 1997). As a supplementary component of the CSCCM, discussions provide students with a direct means of sharing, negotiating, and integrating their ideas to externalize their internal knowledge through textual or oral communications (Ifenthaler et al., 2011). The data generated from students' textual or oral communications provide analytical possibilities for understanding and analysis of students' knowledge. In summary, extracting knowledge from student discussions can be considered as a crucial means to gain insights into the students' knowledge construction process.

Existing knowledge extraction, classification, and analysis methods

Existing research has utilized manual coding and semi-automatic analysis to extract and analyze knowledge. On the one hand, knowledge classification frameworks were used to manually identify students' domain-specific knowledge types and knowledge depths, such as Structure of Observed Learning Outcome (Liu et al., 2021), Revised Bloom's Taxonomy (Blooma et al., 2013), and Technological Pedagogical Content Knowledge (Phillips et al., 2019). On the other hand, semi-automatic analysis relies on AI algorithms to train knowledge classification models based on a labeled training corpus that contains different types of knowledge. Typical AI algorithms include Support Vector Machine (SVM) (Karlovčec et al., 2012), Artificial Neural Network (ANN) (Patikorn et al., 2019), and Bidirectional Encoder Representations from Transformers (BERT) (Shen et al., 2021). Of these two methods, manual coding students' knowledge is labor-intensive, time-consuming, and error-prone work (Han et al., 2021), while the semi-automatic analysis approaches can increase the efficiency of the work, reduce the potential for bias or errors, and improve the reproducibility of data analysis. However, overfitting existed in semi-automatic classification through validation of existing models, resulting in a low accuracy and generalizability when applied to new datasets (Patikorn et al., 2019; Shen et al., 2021). To address these challenges, recent studies have explored automatic analysis methods that can overcome these limitations and provide more accurate and reliable results. Compared to manual coding and semi-automatic analysis, the automatic analysis does not rely on human intervention. Instead, it utilizes advanced technologies to automatically learn and infer knowledge types from unlabeled data. One promising approach to automatic analysis is the use of semantic knowledge analysis, which involves the application of natural language processing and machine learning techniques to extract meanings and relationships from textual data (Liu & Chen, 2023; Pfiffner, 2021; Yeari & van den Broek, 2016).

Semantic knowledge analysis

Semantic knowledge is a set of concepts extracted from linguistic units (i.e., words, sentences, chapters) generated from natural languages or texts (Lupyan et al., 2019). A major approach of semantic knowledge analysis requires grouping or clustering of keywords with methods such as semantic network analysis, topic modeling, and then defining knowledge types based on the meaning of keywords (Drieger, 2013; Gurcan & Cagiltay, 2019; Peng & Xu, 2020). For instance, Drieger (2013) measured node-based clustering coefficients of semantic networks and obtained various local clusters that encoded different semantic knowledge. Gurcan and Cagiltay (2019) used Latent Dirichlet Allocation (LDA) to discover the knowledge domains and skill sets from a textual corpus related to big data software engineering discipline and results extracted ten core competency areas from 48 trending knowledge. However, these methods involve the researchers' manual definition of the knowledge types according to the meaning of the keywords, which may increase workload, lessen the efficiency, and reduce the interpretability of results (Jelodar et al., 2019). To solve this issue, semantic dictionaries or knowledge bases are designed, which apply natural language processing to provide solutions for classifying knowledge based on keywords. Sememe knowledge base is a type of semantic knowledge base that utilizes sememes for describing and organizing the meaning of words and phrases (Zhao et al., 2022). Sememes are defined as the minimum semantic units of human languages in linguistics (Bloomfield, 1926) and a limited set of sememes compose the meanings of all the words. For example, the sememe of "apple" includes *Computer* and *Fruit*, which means the word "apple" has two main meanings: one is a famous computer brand (Apple brand) and another is a sort of juicy fruit (apple). Most research designed and applied sememe knowledge base in the natural language processing, information retrieval, and machine translation fields in order to enhance the computer's ability to understand human language (Niu et al., 2017; Wen et al., 2022; Ye et al., 2022). Recently, some studies have applied sememe to support instruction and learning in the education field. For instance, Liu et al. (2018) developed a mixed similarity strategy to integrate sememe knowledge, orthographic, and phonological features for the automated generation of questions, thereby helping instructors save time in constructing examination papers. Chen and Dong (2022) designed an automatic grading system with text similarity based on sememe to score subjective items in examinations. Given the promises of applying sememe in the education field, it is necessary to investigate how sememe can be used to address educational challenges, such as understanding students' knowledge construction process.

In addition, recent studies have started to further analyze knowledge after extracting and classifying knowledge in order to obtain a comprehensive understanding of students' knowledge construction and advancement (Blooma et al., 2013; Lin et al., 2013; Zhang et al., 2019). These studies focused on a single analytical perspective to analyze knowledge features, including the frequency of certain knowledge types and knowledge structures. For example, Blooma et al. (2013) counted the types of knowledge to capture knowledge characteristics in the CKC process. Results found that "procedural knowledge" was the most prominent knowledge while "meta-cognitive knowledge" was lacking. Zhang et al. (2019) used epistemic network analysis, a learning analytics method, to compare the epistemic network characteristics of teachers' knowledge in different groups. Results found that the teachers with higher scores had a richer, more organized, and more flexible knowledge structure than teachers with lower scores. Although they made valuable attempts to analyze students' knowledge, they often fall short of providing a deep insight into the complexity and dynamics of knowledge construction and development. On the one hand, knowledge has a hierarchical structure and organizational form, which requires a systematic method for classification, integration, and analysis (Daft & Lewin, 1993). On the other hand, knowledge is a dynamic concept constantly develops and evolves over time during students' learning process (Nonaka et al., 2000). Due to the complexity and dynamics of knowledge, merely focusing on one analytical perspective may cause inconclusive and incomprehensible results. An integrated approach enables researchers to gain a comprehensive understanding of complex phenomena, avoid inconclusive or incomplete results, and leverage multiple perspectives to develop effective solutions (Kelley & Knowles, 2016; Sun et al., 2021). Considering the complex and dynamic characteristics of collaborative knowledge construction, it is necessary to use an integrated approach to analyze and understand students' knowledge characteristics and evolutions.

Methodology

Research purposes and questions

This research's purpose was to gain a deep understanding of the students' knowledge characteristics and evolutions during the CKC process by using automatic knowledge analysis methods. This research conducted a series of CSCCM activities supported with online discussions in an *online collaborative concept mapping platform* designed by the research team to facilitate higher education students' CKC quality. Then, this research aimed to extract students' knowledge generated in online discussions and compare semantic knowledge characteristics and evolutionary trends between pairs with high and low performances. There were two research questions:

RQ 1 What were the differences in semantic knowledge characteristics between pairs with high and low performances during the CSCCM process?

RQ 2 What were the differences in evolutionary trends of semantic knowledge between pairs with high and low performances during the CSCCM process?

Research context and participants

The research context was a four-day graduate-level online course titled "*Educational Technology Development and Application*" during summer 2022, offered at a top China's research-intensive university. This course focused on learning theories, instructional practices, and technology applications, as well as development trends in educational technology.

Participants were 16 (15 females and 1 male; ages between 24 to 32) part-time Master of Education students from the College of Education at the university. They came from the majors of educational management (10 students), subject education (5 students), and educational technology (1 student). Participants were divided into 8 pairs randomly. The reasons for choosing participants in this course are twofold: first, the course has been designed with the application of collaborative knowledge construction pedagogy, which is suitable for the research purposes; second, the instruction and learning strategy of concept mapping can be easily learned by the participants, as it has been successfully implemented within higher education. All participants signed the informed consent forms before the course and agreed the data collection of this research.

The instructional process

The online course lasted four days, with the following instructional process designed (see Fig. 1). The first day was designed as an initiation and warm-up phase for students to get familiar with the experiment, concept mapping, and the platform environment. The current experiments were held from Day 2 to Day 4. Each day consisted of three sessions: the online lecture, the individual-level concept mapping activity, and the CSCCM activity. The online lectures included four themes, namely the development of educational technology, online and blended learning, learning analytics and educational data mining, and artificial intelligence in education. The individual-level concept map activity

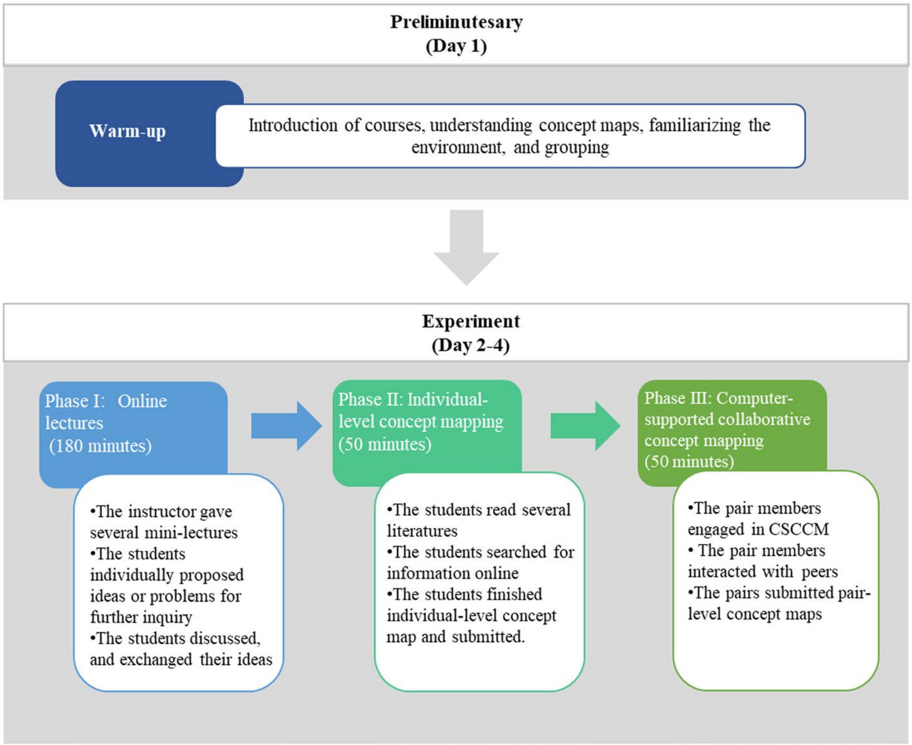
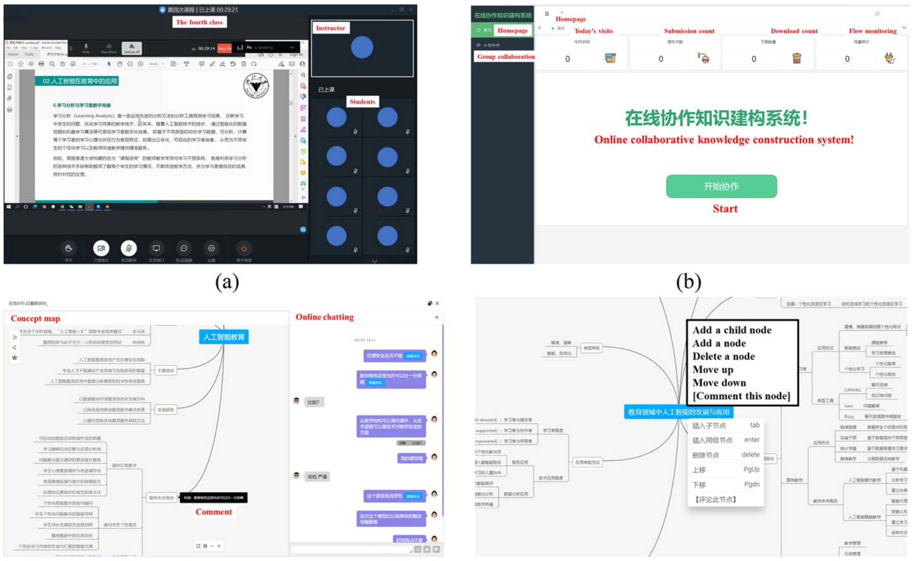

Preliminutesary
(Day 1)
Warm-up
Introduction of courses, understanding concept maps, familiarizing the environment, and grouping
Experiment
(Day 2-4)
Phase I: Online lectures (180 minutes)
•The instructor gave several mini-lectures
•The students individually proposed ideas or problems for further inquiry
•The students discussed, and exchanged their ideas
Phase II: Individual-level concept mapping (50 minutes)
•The students read several literatures
•The students searched for information online
•The students finished individual-level concept map and submitted.
Phase III: Computer-supported collaborative concept mapping (50 minutes)
•The pair members engaged in CSCCM
• The pair members interacted with peers
•The pairs submitted pair-level concept maps

Fig. 1 The instructional procedure

was designed as a preparation for CSCCM (de Weerd et al., 2017). Each student was required to search for learning materials and resources and build an individual-level concept map independently. Then, CSCCM involved student pairs to collaboratively complete a concept map at the pair level. For example, one CSCCM activity asked pairs to record the basic concepts, definitions, theories, instructional processes, and technical support related to blended learning.

The platforms used in this research included DingTalk (see Fig. 2a) and an *online collaborative concept mapping platform* (see Fig. 2b). DingTalk was used for lecturing and student pair's communication during CSCCM activities. *Online collaborative concept mapping platform* was designed by our research team to support individual-level and pair-level concept mapping (see Fig. 2b). The administrator created individual spaces for each student and pair spaces for each pair in advance, so that students can complete their individual-level concept maps and pair-level concept maps. An online chat box was embedded to help students share and exchange their ideas and knowledge while constructing pair-level concept maps (see Fig. 2c). In addition, the platform supports peer evaluation function, which enables students to examine and evaluate peers' concept maps (see Fig. 2d).

On Days 2 – 4, students were required to construct individual-level concept maps and then pair-level concept maps after the lecture. In the CSCCM activity on Day 2, students were asked to engage in the CSCCM activity immediately after completing the individual-level concept maps. In the CSCCM activity on Day 3, students were asked to view peers' concept maps before constructing pair-level concept maps for cognitive group awareness support (Farrokhnia et al., 2019). Furthermore, in the CSCCM activity on Day 4, students had to evaluate peers' concept maps, respond to peers' comments, and modify individual-level concept maps before constructing pair-level concept maps to improve the completeness of their concept mapping (Hwang & Chang, 2021). The



(a) (b)

(c) (d)

Fig. 2 Screenshots of **a** DingTalk, **b** online collaborative concept mapping platform, **c** online chatting function in the platform, and **d** online commenting function in the platform

CSCCM activities included dialogic prompts to foster knowledge inquiry and construction (e.g., What do you think about this idea?, Do you agree with my ideas?, My opinion is ..., I disagree with this idea because..., The idea about ...is appropriate, A summary of our pair's idea is...).

Data collection and analysis process

This research collected data in two ways. First, discussion data from eight pairs on the *online collaborative concept mapping platform* and DingTalk during the CSCCM activities were collected, mainly including discussion content, discussion participants, and discussion time. There was a total of 681 discussion data. Second, the final versions of the pair-level concept maps were collected; there was a total of 24 concept maps (8 pairs * 3 times). An overall analytical framework was proposed, which used an integrated approach to analyze semantic knowledge characteristics and evolutionary trends between the pairs with high and low performances (see Fig. 3).

Regarding the pair-level concept maps, an assessment standard was adapted to evaluate pair-level concept maps (see Table 1). The pair-level concept maps were evaluated in three dimensions: structure (distribution of nodes), idea (average depth of ideas), and connection (average depth of connections). The overall score was obtained by adding the scores of distributions of nodes (DIS), average depth of ideas (DoI), and average depth of connections (DoC). The performance for each pair was defined as the average score of the pair-level concept maps completed for the three CSCCM activities. Two raters with educational technology background calculated the three values for the pair-level concept maps

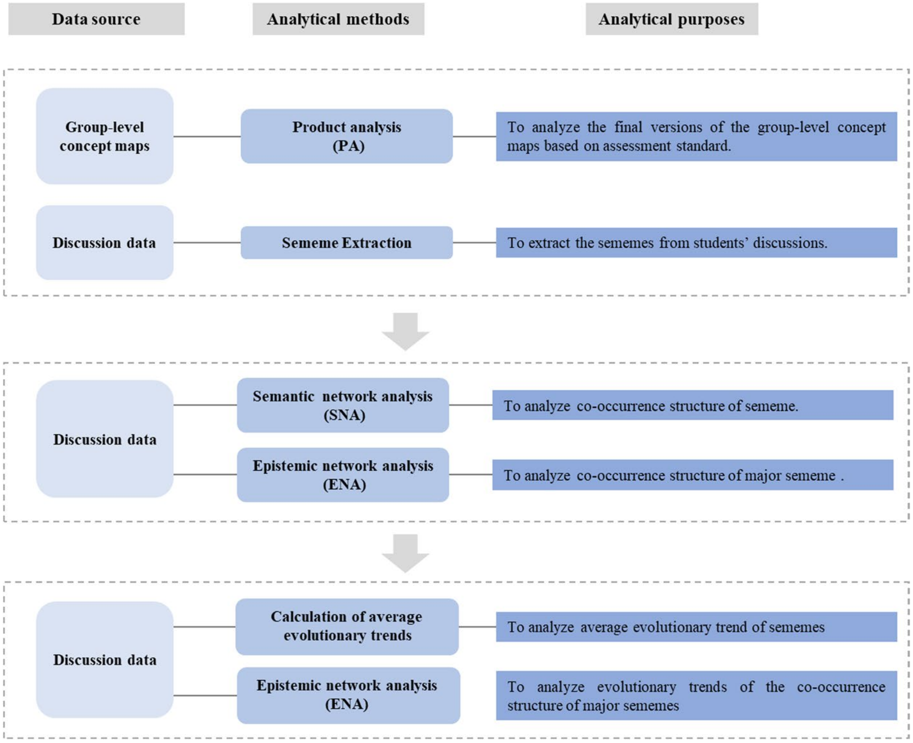

Data source
Analytical methods
Analytical purposes
Group-level concept maps
Product analysis (PA)
To analyze the final versions of the group-level concept maps based on assessment standard.
Discussion data
Sememe Extraction
To extract the sememes from students' discussions.
Discussion data
Semantic network analysis (SNA)
To analyze co-occurrence structure of sememe.
Epistemic network analysis (ENA)
To analyze co-occurrence structure of major sememe .
Discussion data
Calculation of average evolutionary trends
To analyze average evolutionary trend of sememes
Epistemic network analysis (ENA)
To analyze evolutionary trends of the co-occurrence structure of major sememes

Fig. 3 The analytical framework

Table 1 Assessment standard for pair-level concept maps

Dimension	Indicator	Code	Description
Structure	Number of layers	NoL	The number of layers in each concept map
	Number of nodes	NoN	The number of nodes in each concept map
	Number of connections	NoC	The number of connections in each concept map
	Distributions of nodes	DIS	The score of distribution was measured as the quantity of ideas divided by the number of layers (DIS = NoN / NoL). A higher score of DIS indicated a larger number of nodes distributed for each hierarchy
Idea	Overall quality of ideas	QoI	The weighted sum of the number of nodes at each level (QoI = level 1 * 1 + level 2 * 1.5 + level 3 * 2 + level 4 * 2.5 + …). A higher score of QoI indicated a higher quality of a concept map
	Average depth of ideas	DoI	DoI was measured as the overall quality of ideas divided by the number of nodes (DoI = QoI/NoN). A higher score of DoI indicated a deeper idea elaboration on average
Connection	Overall quality of connections	QoC	The weighted sum of the number of connections for each type (QoC = a * 1 + b * 2) a: There is no relation between two nodes; b: There is an interpretable relation between two nodes
	Average depth of connections	DoC	DoC was measured as the Overall quality of connections divided by the number of connections (DoC = QoC/NoC)

independently, and they reached agreements through discussions when there were conflicts about the scoring. According to the scoring results, eight pairs were divided into high-performing pairs and low-performing pairs. The high-performing pairs consisting of four pairs (i.e., pair 2, 3, 4, and 7) achieved higher scores for their pair-level concept maps (M = 22.72, SD = 1.10), while the low-performing pairs, also consisting of four pairs (i.e., pair 1, 5, 6, and 8) achieved lower scores for their pair-level concept maps (M = 17.48, SD = 1.10).

The top 100 keywords were extracted from the discussion data and sememes that correspond to keywords were identified. One keyword may belong to several different sememes and we decided to choose a sememe that occurred frequently as the sememe for this keyword. For example, the sememe of a keyword "theory" includes *Debate* and *Knowledge*, and the sememe of a keyword "misconception" includes *Wrong* and *Knowledge*. The sememe *Knowledge* appeared with a high frequency, therefore the sememe of keywords "theory" and "misconception" was defined as *Knowledge* in this research. Finally, the top 100 keywords were identified as 65 sememes. 60 and 52 sememes were identified in the high-performing pairs and low-performing pairs, respectively.

To answer the first question, semantic network analysis (SNA) and epistemic network analysis (ENA) were used to compare sememe characteristics between the high-performing pairs and low-performing pairs. First, the sememe networks of high-performing and low-performing pairs were created to identify co-occurrence structures using the network visualization software, Gephi. Pointwise Mutual Information (PMI) was used to measure the probability of two sememes appearing in all discussions. PMI was represented as

$$PMI(S_1, S_2) = \log(\frac{P(S_1, S_2)}{P(S_1)P(S_2)})$$

where $P(S_1, S_2)$ represented the probability of the co-occurrence of S_1 and S_2, $P(S_1)$ represented the probability of S_1 occurrence, and $P(S_2)$ represented the probability of S_2

Table 2 The descriptions of SNA metrics

SNA metrics	Description
Density	The ratio of actual links between any nodes to all potential possible links
Average Path Length (APL)	The average number of the shortest paths for all possible pairs of nodes
Transitivity	The ratio of transitive triads to all triads in a network
Reciprocity	The ratio of symmetric dyads to non-null dyads in a network
Centralization	The extent to which degree centrality is concentrated in a network ranged between 0 and 1
Distance	The maximum value of the distance between any two nodes in the network
Average Degree (AD)	The mean of all nodes' degrees
Average Weighting Degree (AWD)	The mean of all nodes' weighted degree

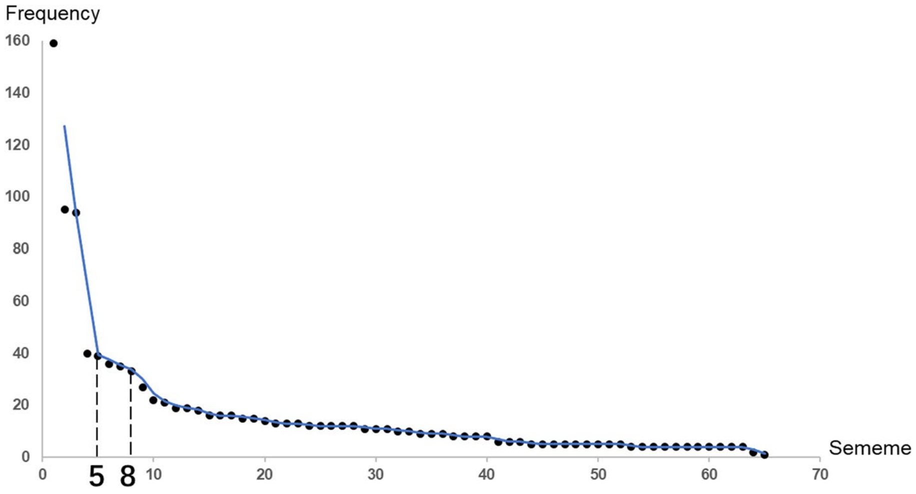


Fig. 4 A frequency scatter plot for sememe

occurrence. Modularity analysis, a community detection method based on the Leuven algorithm, was conducted to reveal different clusters within a sememe network. The pair-level SNA metrics were calculated to uncover semantic network characteristics, including density, average path length (APL), transitivity, reciprocity, centralization, distance, and average weighting degree (AWD) (see Table 2) (Ouyang et al., 2021). R package *igraph* was used to measure those SNA metrics.

Moreover, ENA was used to analyze the co-occurrence structure of major sememes (i.e., sememes with high frequencies). A scatter plot about sememe frequencies was drawn to determine the number of sememes for ENA (see Fig. 4). Sememes were arranged on the scatter plot in descending order of the frequencies. The trend line of the scatter plot showed a sudden change when the number of sememes was 5 or 8. In order to display the co-occurrence relationship between sememes as much as possible, we chose 8 sememes as the ENA nodes, which were *Education*, *Image*, *Knowledge*, *Study*, *Plans*, *NounUnit*, *Thinking*, and *FuncWord* (see Table 3).

Note. X-axis represented sememes sorted from low to high frequency. Y-axis represented the frequencies of sememes.

To answer the second research question, we used 10 min as a time slice to analyze the evolutionary trends of sememes, and each CSCCM activity was divided into 5 stages

Table 3 Sememes selected for ENA

Rank	Sememe	Frequency
1	*Education*/教育	159
2	*Image*/图像	95
3	*Knowledge*/知识	94
4	*Study*/学习	40
5	*Plans*/规划	39
6	*NounUnit*/名量	36
7	*Thinking*/思想	35
8	*FuncWord*/功能词	33

with a total of 50 min. Firstly, we constructed time series of sememes in this research. It can be defined as

$$k_{i,j} = \frac{P_{i,j}}{\sum_{j=1}^{n} P_{i,j}} \left(i = 1, 2, \ldots, m; j = 1, 2, \ldots, 5\right)$$

where $k_{i,j}$ represented the relative frequency of the i-th sememe on the j-th 10-min and $P_{i,j}$ represented the frequency of the i-th sememe on the j-th 10-min. Therefore, the time series matrix of sememes can be represented as

$$K = [K_1, K_2, \cdots, K_m]^T = \begin{bmatrix} k_{1,1} & k_{1,2} & \cdots & k_{1,5} \\ k_{2,1} & k_{2,2} & \cdots & k_{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m,1} & k_{m,2} & \cdots & k_{m,5} \end{bmatrix}$$

According to the results of modularity analysis, we calculated the average evolutionary trend of each cluster in the high-performing and low-performing pairs based on Euclidean distance. Euclidean distance is a commonly used definition of distance that calculates the arithmetic mean value of each time slice (Aghabozorgi et al., 2015). The average evolutionary trend was considered as the overall characteristic of each cluster. Moreover, ENA was performed to characterize evolution of major sememes for the five stages of CSCCM activities in the high-performing and low-performing pairs.

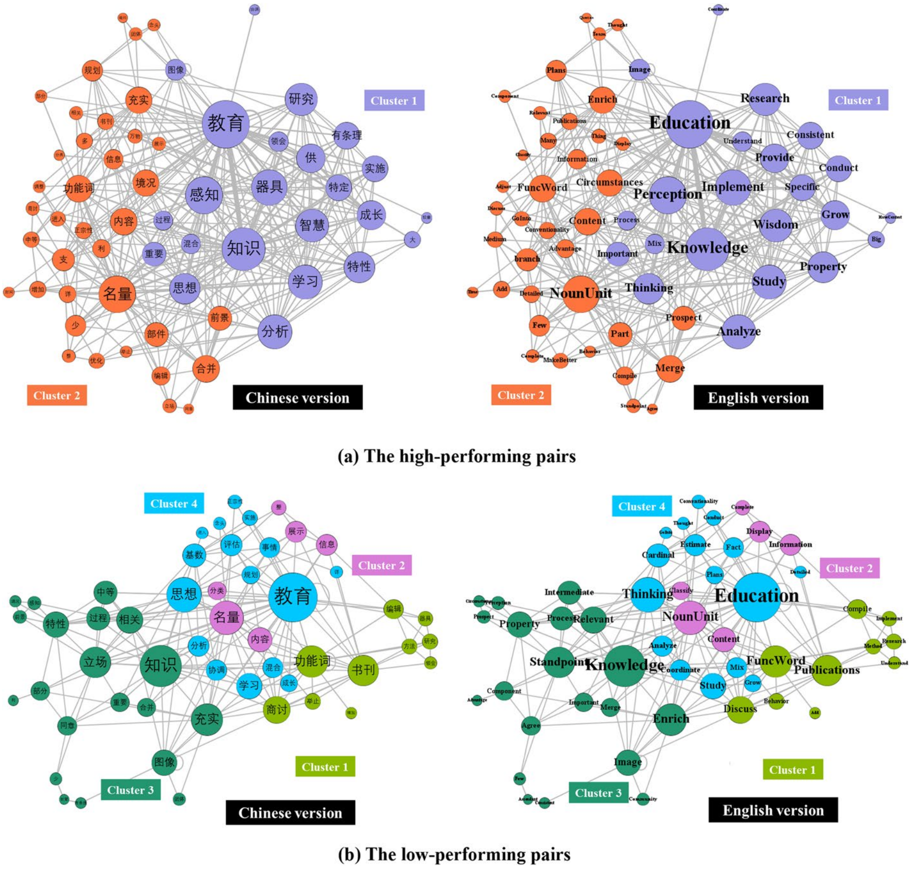
Results

RQ 1: What were the differences in semantic knowledge characteristics between pairs with high and low performances during CSCCM?

Regarding the semantic network analysis results, the high-performing pairs demonstrated higher connectedness and stronger cohesion than the low-performing pairs. Specifically, high-performing pairs had higher values of density, transitivity, centralization, average degree, and average weighting degree than low-performing pairs. In addition, high-performing pairs had lower values of average path length and distance than low-performing pairs (see Table 4). In summary, the high-performing pairs formed a semantic network with high connectedness and strong cohesion while the low-performing pairs formed a semantic network with low connectedness and weak cohesion.

Table 4 The comparison of the SNA metrics of high-performance and low-performing pairs

SNA metrics	Density	APL	Transitivity	Reciprocity	Centralization	Distance	AD	AWD
High-performing	0.335	1.898	0.476	1	2.512	4	19.767	36.016
Low-performing	0.176	2.256	0.387	1	0.745	5	9.0	17.225



(a) The high-performing pairs

(b) The low-performing pairs

Fig. 5 Sememes network diagrams

Modularity analysis generated two clusters for the high-performing pairs and four clusters for the low-performing pairs (see Fig. 5). For the high-performing pairs, cluster 1 centered on the education-related theory and practice, technology applications and developments in education, and cluster 2 centered on grammatical meanings, forms, and functions. Specifically, approximately 38.33% of sememe were clustered in cluster 1. The core sememes in cluster 1 were *Knowledge*, *Education*, *Implement*, *Perception,* and *Study.* Therefore, cluster 1 was about course lectures and CSCCM specific content. Cluster 2 consisted of 61.67% of sememes without full lexical meanings but with grammatical meanings and grammatical functions, such as *FuncWord* and *NounUnit*. In addition, *Enrich* and *Merge* indicated that students paid attention to the adjustment and modification of the concept map. For low-performing pairs, cluster 1 and cluster 2 centered on grammatical meanings, forms, and functions. Cluster 3 and cluster 4 centered on education-related content and topic. Specifically, the core sememes in each of the four clusters were *FuncWord*, *NounUnit*, *Knowledge*, and *Education*, respectively. These clusters covered 18.87%, 13.21%, 37.74%, and 30.19% of the sememe frequency. In summary, sememes were more tightly connected in the high-performing pairs than in the low-performing pairs. Moreover, sememes related to the learning content were clustered into one cluster in the high-performing pairs, which was reflected by a strong sememe connection in cluster 1. However, sememes related to the learning content were divided into two different clusters in the low-performing pairs, including cluster 3 and cluster 4, which meant that students did not integrate a variety of knowledge related to the learning content in their discussions.

Note. Nodes represented sememes and nodes in different colors represented different clusters. Node size represented relative influence, i.e., eigenvector centrality. Tie weights represented the strength of relations, i.e., co-occurrence frequency of two sememes.

ENA results showed the co-occurrence structure of major sememes between the high-performing pairs and low-performing pairs. The high-performing pairs and low-performing pairs were characterized by the connection values and the locations of the centroid of the ENA plots (see Fig. 6). For all pairs, most of the codes shared strong connections with *Education*, the core theme of the course content. However, the sememe connected to *Education* was completely different, which can be reflected by the locations of the centroid in epistemic networks. Specially, for the high-performing pairs, the centroid of the epistemic network was located to the left of X-axis, mainly focusing on *Knowledge*, *Plans, Study*, and *Education*. The connection between *Education* and *Knowledge* was 0.43; the connection between *Education* and *Plans* was 0.33; and the connection between *Education* and *Study* was 0.17. For the low-performing pairs, the centroid of the epistemic network was located on the positive axis for X, focusing on *NounUnit*, *FuncWord*, *Image*, and *Education*. The connection between *Education* and *NounUnit* was 0.38; the connection between *Education* and *Image* was 0.28; and the connection between *Education* and *NounUnit* was 0.19. Moreover, Mann–Whitney U test further revealed the differences in the distribution of connection between the high-performing

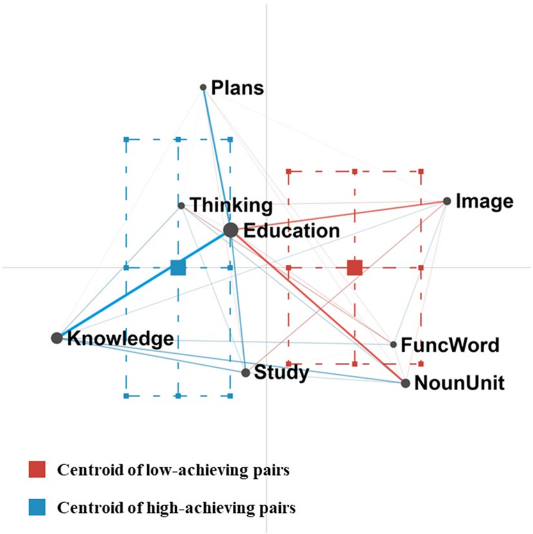

Plans
Thinking
Education
Image
Knowledge
FuncWord
Study
NounUnit
🟥 Centroid of low-achieving pairs
🟦 Centroid of high-achieving pairs

Fig. 6 The subtracted ENA plots of the high- and low-performing pairs

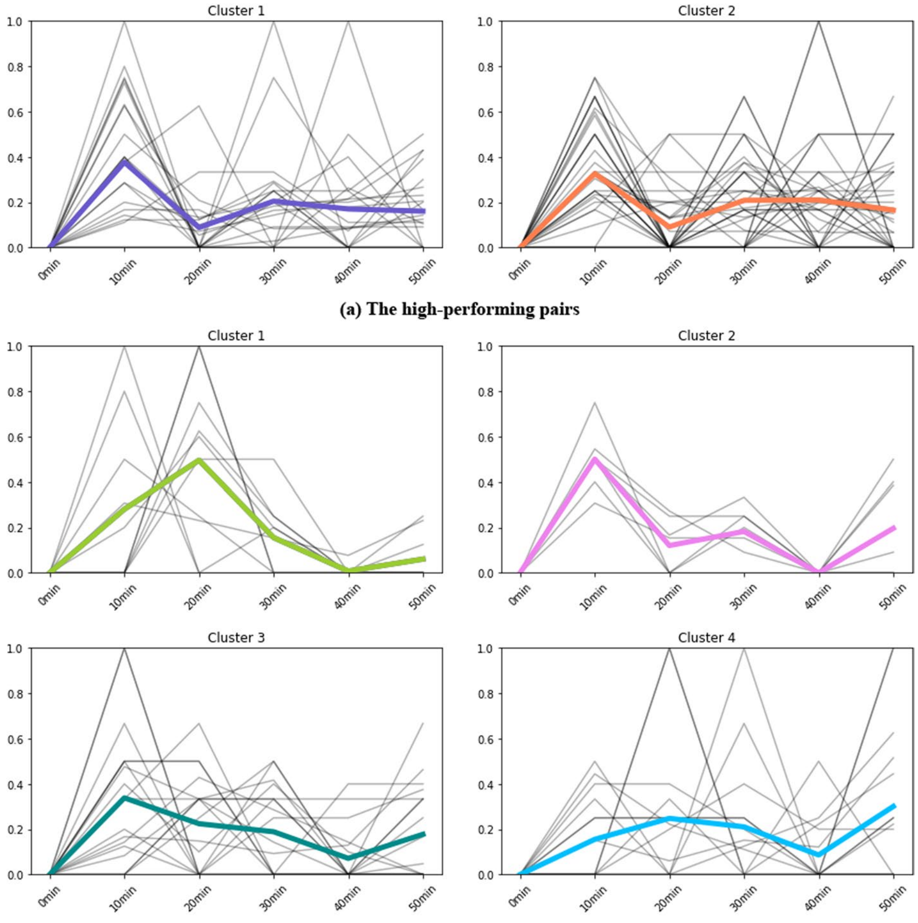
pairs and low-performing pairs. A significant difference was found on the X-axis (U = 60, p = 0.00, r = − 0.87), which meant that there were significant differences in the connections between the high-performing pairs and low-performing pairs. In summary, the high-performing pairs concentrated on discussing content related to the course content and CSCCM activities; the low-performing pairs concentrated on discussing grammatical meanings and functions.

Note. In subtracted network, the blue square represented the centroid of high-performing pairs, the red square represented the centroid of low-performing pairs and the boxes represented 95% confidence intervals. The weights of the connections were compared between high-performing pairs and low-performing pairs, and the color of the line was set to be the same as the pair that had a stronger connection between the sememe. The color depth represents the strength of the connection.

RQ 2: What were the differences in evolutionary trends of semantic knowledge between pairs with high and low performances during CSCCM?

The evolutionary trends of sememes in the high-performing pairs were relatively stable while sememes in low-performing pairs showed variability and fluctuation. For the high-performing pairs, the average evolutionary trends of sememes in two clusters were roughly similar, as demonstrated by similar evolutionary shapes throughout the CSCCM activities (see Fig. 7a). The range of sememe frequency (i.e., the maximum value of frequency minus the minimum value of frequency) was 0.29 for cluster 1 and 0.24 for cluster 2. In addition, the fluctuation of cluster 1 and cluster 2 occurred during the first half of the activity for the high-performing pairs. For the low-performing pairs, cluster 1 and cluster 2, centering on grammatical meaning, form, and function, had more variabilities and changes, compared to cluster 3 and cluster 4, centering on course-related knowledge (see Fig. 7b). Specifically, for the low-performing pairs, ranges of sememe frequency in four clusters were 0.49, 0.50, 0.27, and 0.21. In addition, the peaks of three clusters (except cluster 4) occurred in the first half of the activity, while valleys of these four clusters occurred in the second half of the activity. In summary, the evolution of sememe in high-performing pairs was relatively stable, compared to the low-performing pairs.

ENA results showed the evolution in the co-occurrence structure of major sememes between the high-performing and low-performing pairs. The high-performing pairs and low-performing pairs were characterized by the locations of the centroid of the ENA plots (see Fig. 8). For high-performing pairs, the centroid for most of the stages fell in the upper half of the network, which indicated that high-performing pairs were able to continuously focus on the course content for knowledge construction during the CSCCM activity. The sememe evolution trend had changes in the middle stage of the activity (i.e., 20–30 min) and the centroid was biased towards *FuncWord.* For the low-performing pairs, the centroid for most of the stages fell in the lower half of the network, which indicated that low-performing pairs were not able to continuously focus on the meaningful learning content during the CSCCM activity. The centroid of low-performing pairs was biased towards course-related sememes, such as *Education* in the middle stage of the activity (i.e., 20–40 min). Moreover, Mann–Whitney U test was used to determine whether there were significant differences in the position of the stage centroid between two adjacent stages. For high-performing pairs, there was no significant difference on



(a) The high-performing pairs

(b) The low-performing pairs

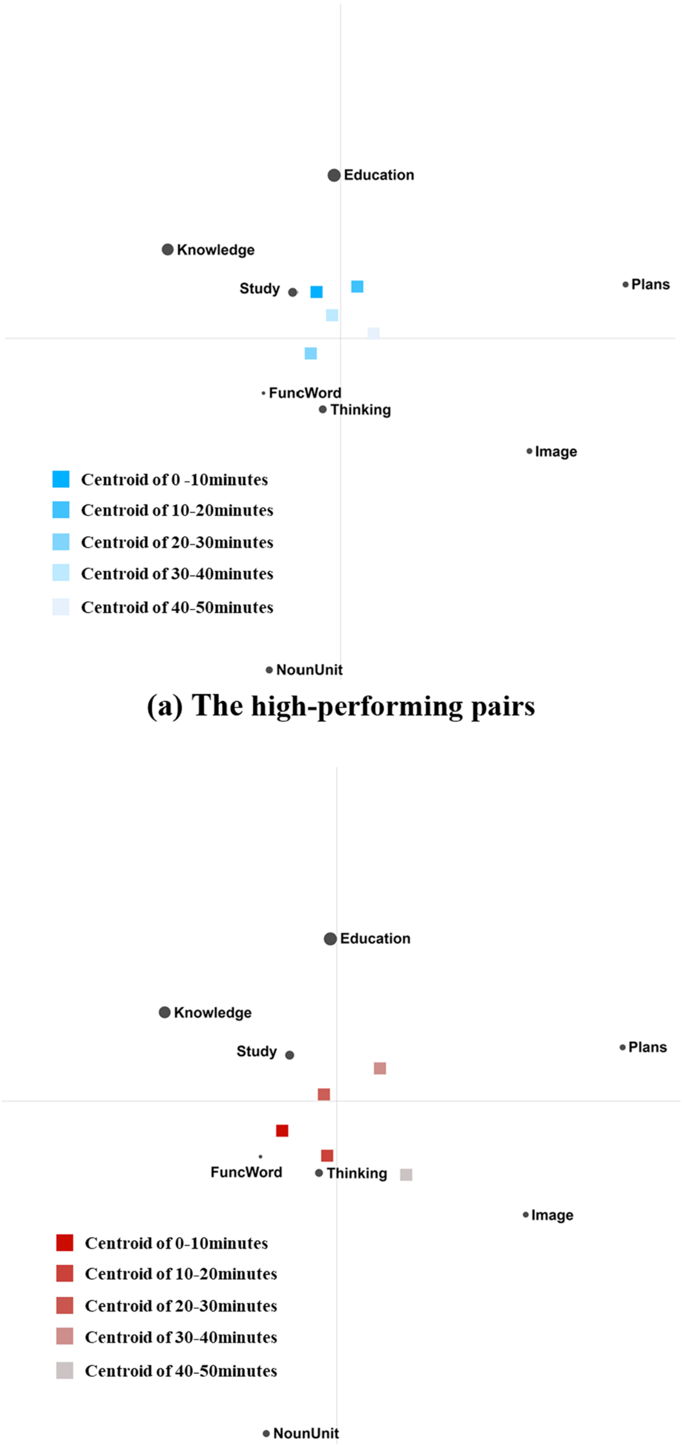
Fig. 7 Average evolutionary trend for clusters in the high- and low-performing pairs. The black line represented the evolution of each sememe and the colored line represented the average evolution of sememe in one cluster.

the X-axis or Y-axis between two adjacent stages. For low-performing pairs, significant differences were found on the Y-axis between the centroid of 10-20 min and the centroid of 20–30 min (U = 10.5, p = 0.02, r = 0.67), and between the centroid of 30–40 min and centroid of 40–50 min (U = 43, p = 0.01, r = − 0. 79). In summary, the co-occurrence structure of the major sememes in high-performing pairs fluctuated slightly, focusing mainly on course-related knowledge, while the co-occurrence structure of the major sememes in low-performing pairs fluctuated greatly, with less attention to course-related knowledge.

Discussions and implications

Addressing the research questions

To gain a deep comprehension of the students' knowledge characteristics and evolutions during the CKC process, this research integrated learning analytics methods with semantic knowledge analysis based on a knowledge base to extract, analyze, and understand students' knowledge construction process. Regarding the first research question, the result showed that high-performing pairs focused on course-related and



(a) The high-performing pairs

(b) The low-performing pairs

Fig. 8 Evolution of the centroid in ENA plots

activity-related knowledge, while the low-performing pairs concentrated on discussing grammatical meanings and functions. Specially, for high-performing pairs, meaningful sememes related to the course content and CSCCM activity (i.e., *Education*, *Learning,* *Knowledge*) formed a strong co-occurrence structure with a high frequency. In addition, meaningful sememes were clustered into one group in the sememe network, indicating that high-performing pairs made full use of learning content and materials and thought comprehensively in the discussion. For low-performing pairs, linguistical knowledge that represents units (i.e., *NounUnit*, *FuncWord*) in language formed a strong co-occurrence structure with a high frequency. In addition, for the low-performing pairs, sememes were not clustered into one group in the sememe network, indicating that students tended to have scattered thoughts and cannot thought comprehensively in the discussion. Overall, the research results showed that semantic knowledge characteristics of high-performing pairs existed a strong focus on course-related and activity-related knowledge while semantic knowledge characteristics of low-performing pairs existed a strong emphasis on linguistic knowledge. Consistent with previous research results (e.g., Peng & Xu, 2020; Yoon et al., 2021), the results indicated that when students concentrate on the content that is relevant to the course, task, and activity during collaborative discussions, they tend to attain good academic results.

Regarding the second research question, the results indicated that the evolutionary trend of sememes in high-performing pairs tended to be relatively stable while low-performing pairs showed variability and fluctuation over time. Specially, compared to low-performing pairs, the high-performing pairs exhibited smaller changes in sememe frequencies throughout the CSCCM activities (reflected by the smaller value of change range in sememe), and lower differences in the co-occurrence structure of major sememes (reflected by closer centroid position). In addition, two clusters in the high-performing pairs had similar evolutionary trends, while four clusters in the low-performing pairs had diverse evolutionary trends. This result again verified that the high-performing pairs had a stable knowledge evolutionary trend but the low-performing pairs showed variability and fluctuation of knowledge evolution. Overall, the research results showed that the high-performing pairs demonstrated a more sustained cognitive engagement, compared to the low-performing pairs (Liu et al., 2022).

Analytical implications

Since CKC is a complex, adaptive, and dynamic process, this research extended the knowledge extraction and analysis using an integrated approach, combining semantic knowledge analysis with learning analytics to gain a comprehensive understanding of knowledge characteristics and evolutions during students' CKC processes. There are two analytical implications generated from this research, namely *the application of domain-specific knowledge bases* and *AI-driven learning analytics and data mining*. First, it is essential to develop semantic dictionaries or knowledge bases that are customized for specific subjects to conduct productive knowledge analysis. Semantic dictionaries or knowledge bases represent all words as a finite set of semantics by defining upper-level semantic properties, which is highly interpretable and the results can be easily understood by students, instructors, and researchers. In general, this research represented the first attempt to apply knowledge bases from the natural language processing domain to knowledge analysis in educational contexts. However, general dictionaries cannot cover all the required terms and concepts in a certain field. Future work can construct, update, and maintain semantic dictionaries or knowledge bases in different specialized domains, thus improving the accuracy of the extracted semantic knowledge. Second, integrated approaches, particularly AI-driven learning analytics and data mining, are worth applying to capture the nature of CKC. Compared to the traditional analytical methods, the integrated approach used in this research, namely integrated learning analytics methods with semantic knowledge analysis can better extract and represent the complex and dynamic structure of CKC (de Carvalho & Zárate, 2020). Future work can apply advanced AI algorithms (e.g., natural language processing and genetic programming) with learning analytics and data mining to offer in-time, dynamic knowledge characteristics of CKC (de Carvalho & Zárate, 2020; Hoppe et al., 2021). For example, Ouyang et al. (2023) proposed an integrated approach that combined a probabilistic model with two sequence analysis and mining techniques to investigate macro-level collaborative patterns and micro-level sequences of group communicative discourses. Overall, to obtain a comprehensive understanding of students' knowledge construction, application of domain-specific knowledge bases and AI-driven learning analytics and data mining have potentials to optimize the process of knowledge extraction and analysis, increase the capacity of generalizability and accuracy of results, and increase the efficiency of the knowledge analysis work.

Pedagogical implications

Three pedagogical implications are proposed based on the semantic knowledge insights generated from our investigation, including reasonable monitoring, appropriate incentive, and adaptive support. First, instructors are supposed to make reasonable monitoring to guide students who are trapped in off-topic discussions in the correct direction. Our results showed that the low-performing pairs focused less on course-related discussions with more fluctuated knowledge evolution, compared to the high-performing pairs. When these students face bottlenecks in the discussion, the instructor should identify areas where they need to improve, encourage them to ask questions and share thoughts, and provide information or hints to promote their innovative thinking (Al-Zahrani, 2015; Golding, 2011). Second, instructors are supposed to promote active involvement and introduce appropriate incentives to encourage students to stay on-topic. Results showed that pairs with low performance wasted time on content without practical meaning, while pairs with high performance focused on content related to course and CSCCM activities. Therefore, instructors should consider ways (such as extra credit or praise) to encourage students to stay focused on the task at hand (Hou & Wu, 2011). Third, instructors are supposed to provide adaptive support in terms of student pairs' and groups' dynamic evolvement in the collaborative learning process. Our results showed considerable fluctuations of knowledge evolution during the initial stages, and a decreased focus of knowledge in the later stages. As mentioned by Ouyang et al. (2023), the collaborative learning process needs a dynamic intervention approach to suit students' complicated tendencies. To be specific, instructors can observe, monitor, and regulate students to organize and focus on discussions about the topic in the first half of the activity, and guide students to inquiry and construct knowledge from multiple perspectives in the second half of the activity. Overall, students' activities should be reasonably monitored, and instructors should introduce appropriate incentives and support students' work appropriately with instructional interventions.

Conclusions, limitations, and future directions

In the era of knowledge-based economy, knowledge has become increasingly extensive, complex, and diverse, which poses high demands on understanding students' knowledge. Using integrated learning analytics methods with semantic knowledge analysis, this research this study offered valuable insights into the complex and dynamic nature of students' knowledge during the CKC process. The results revealed differences between pairs with high and low performances in terms of semantic knowledge characteristics and evolutionary trends. Moreover, this research provided analytical contributions to extract and analyze students' knowledge for the comprehension of student knowledge, and proposed pedagogical implications to advance students' knowledge advancement. There were two limitations of this research, which lead to future research directions. First, the sample size of the research was small and the educational background was homogeneous, which weakened the generalizability of the research results and implications. Future research should expand the sample size to different instructional contexts in order to verify the research results and implications. Second, this research mainly employed SNA and ENA to uncover the knowledge characteristics and evolution trends of students, which may not be sufficient for a comprehensive analysis of knowledge. Further research can integrate additional analytical methods (such as lag sequential analysis) and AI algorithms (such as probabilistic model) to provide a more complete description of knowledge construction process. Overall, this research provided researchers and educators with new insights into complex and dunamic nature of CKC and offered analytical implications for understanding students' knowledge in a comprehensive way, which is essential for educational practice and research in the knowledge era.

Acknowledgements

We thank the students participated in this research.

Author contributions

Ning Zhang: Collection and analysis of research data, and writing of the original draft. Fan Ouyang: Conceptualization and manuscript writing (review, editing), supervision of the research process. All authors read and approved the final manuscript.

Funding

The research is supported by National Natural Science Foundation of China (62177041), and the Fundamental Research Funds for the Central Universities, Zhejiang University, China.

Availability of data and materials

The data will be available on request from the corresponding author.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 1 June 2023 Accepted: 11 July 2023
Published online: 01 August 2023

References

Aghabozorgi, S., SeyedShirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering—A decade review. *Information Systems, 53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Al-Zahrani, A. M. (2015). From passive to active: The impact of the flipped classroom through social learning platforms on higher education students' creative thinking. *British Journal of Educational Technology, 46*(6), 1133–1148. https://doi.org/10.1111/bjet.12353

Anderson, R. E. (2008). Implications of the information and knowledge society for education. In J. Voogt & G. Knezek (Eds.), *International Handbook of Information Technology in Primary and Secondary Education* (pp. 5–22). Springer US. https://doi.org/10.1007/978-0-387-73315-9_1

Ashwin, P. (2014). Knowledge, curriculum and student understanding in higher education. *Higher Education, 67*(2), 123–126. https://doi.org/10.1007/s10734-014-9715-3

Blooma, M. J., Kurian, J. C., Chua, A. Y. K., Goh, D. H. L., & Lien, N. H. (2013). Social question answering: Analyzing knowledge, cognitive processes and social dimensions of micro-collaborations. *Computers & Education, 69*, 109–120. https://doi.org/10.1016/j.compedu.2013.07.006

Bloomfield, L. (1926). A set of postulates for the science of language. *Language, 2*(3), 153–164. https://doi.org/10.2307/408741

Chang, C.-C., Liu, G.-Y., Chen, K.-J., Huang, C.-H., Lai, Y.-M., & Yeh, T.-K. (2017). The effects of a collaborative computer-based concept mapping strategy on geographic science performance in junior high school students. *Eurasia Journal of Mathematics, Science and Technology Education, 13*(8), 5049–5060. https://doi.org/10.12973/eurasia.2017.00981a

Charlton, P., & Avramides, K. (2016). Knowledge construction in computer science and engineering when learning through making. *IEEE Transactions on Learning Technologies, 9*(4), 379–390. https://doi.org/10.1109/TLT.2016.2627567

Chen, M., & Dong, Y. (2022). Design of exercise grading system based on text similarity computing. *Mobile Information Systems, 2022*, 1–7. https://doi.org/10.1155/2022/4634903

Chiou, C. (2008). The effect of concept mapping on students' learning achievements and interests. *Innovations in Education and Teaching International, 45*(4), 375–387. https://doi.org/10.1080/14703290802377240

Chu, H.-C., Wang, C.-C., & Wang, L. (2019). Impacts of concept map-based collaborative mobile gaming on English grammar learning performance and behaviors. *Journal of Educational Technology & Society*, *22*(2), 86–100. Retrieved from https://www.jstor.org/stable/26819619

Daft, R. L., & Lewin, A. Y. (1993). Where are the theories for the "New" organizational forms? An editorial essay. *Organization Science*, *4*(4), i–vi. Retrieved from https://www.jstor.org/stable/2635077

de Carvalho, W. F., & Zárate, L. E. (2020). A new local causal learning algorithm applied in learning analytics. *The International Journal of Information and Learning Technology, 38*(1), 103–115. https://doi.org/10.1108/IJILT-04-2020-0046

Drieger, P. (2013). Semantic network analysis as a method for visual text analytics. *Procedia - Social and Behavioral Sciences, 79*, 4–17. https://doi.org/10.1016/j.sbspro.2013.05.053

Durst, S., & Zieba, M. (2018). Mapping knowledge risks: Towards a better understanding of knowledge management. *Knowledge Management Research & Practice., 17*(1), 1–13. https://doi.org/10.1080/14778238.2018.1538603

Farrokhnia, M., Pijeira-Díaz, H. J., Noroozi, O., & Hatami, J. (2019). Computer-supported collaborative concept mapping: The effects of different instructional designs on conceptual understanding and knowledge co-construction. *Computers & Education, 142*, 1–15. https://doi.org/10.1016/j.compedu.2019.103640

Felder, R. M., & Brent, R. (2005). Understanding student differences. *Journal of Engineering Education, 94*(1), 57–72. https://doi.org/10.1002/j.2168-9830.2005.tb00829.x

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction, 12*(2), 213–232. https://doi.org/10.1016/S0959-4752(01)00005-6

Golding, C. (2011). Educating for critical thinking: Thought-encouraging questions in a community of inquiry. *Higher Education Research and Development, 30*(3), 357–370. https://doi.org/10.1080/07294360.2010.499144

Greene, J. A., & Azeved, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist, 45*(4), 203–209. https://doi.org/10.1080/00461520.2010.515935

Gurcan, F., & Cagiltay, N. E. (2019). Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access, 7*, 82541–82552. https://doi.org/10.1109/ACCESS.2019.2924075

Han, Z. M., Huang, C. Q., Yu, J. H., & Tsai, C. C. (2021). Identifying patterns of epistemic emotions with respect to interactions in massive online open courses using deep learning and social network analysis. *Computers in Human Behavior, 122*, 1–16. https://doi.org/10.1016/j.chb.2021.106843

Hoppe, H. U., Doberstein, D., & Hecking, T. (2021). Using sequence analysis to determine the well-functioning of small groups in large online courses. *International Journal of Artificial Intelligence in Education, 31*(4), 680–699. https://doi.org/10.1007/s40593-020-00229-9

Hou, H.-T., & Wu, S.-Y. (2011). Analyzing the social knowledge construction behavioral patterns of an online synchronous collaborative discussion instructional activity using an instant messaging tool: A case study. *Computers & Education, 57*(2), 1459–1468. https://doi.org/10.1016/j.compedu.2011.02.012

Hwang, G.-J., & Chang, S.-C. (2021). Facilitating knowledge construction in mobile learning contexts: A bi-directional peer-assessment approach. *British Journal of Educational Technology, 52*(1), 337–357. https://doi.org/10.1111/bjet.13001

Ifenthaler, D., Masduki, I., & Seel, N. M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. *Instructional Science, 39*(1), 41–61. https://doi.org/10.1007/s11251-009-9097-6

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development, 45*(1), 65–94. https://doi.org/10.1007/BF02299613

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A. (2012). Knowledge component suggestion for untagged content in an intelligent tutoring system. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 195–200). Springer. https://doi.org/10.1007/978-3-642-30950-2_25

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education, 3*(1), 1–11. https://doi.org/10.1186/s40594-016-0046-z

Lehmann, T., Hähnlein, I., & Ifenthaler, D. (2014). Cognitive, metacognitive and motivational perspectives on preflection in self-regulated online learning. *Computers in Human Behavior, 32*, 313–323. https://doi.org/10.1016/j.chb.2013.07.051

Lin, P.-C., Hou, H.-T., Wang, S.-M., & Chang, K.-E. (2013). Analyzing knowledge dimensions and cognitive process of a project-based online discussion instructional activity using Facebook in an adult and continuing education course. *Computers & Education, 60*(1), 110–121. https://doi.org/10.1016/j.compedu.2012.07.017

Liu, M., Rus, V., & Liu, L. (2018). Automatic Chinese multiple choice question generation using mixed similarity strategy. *IEEE Transactions on Learning Technologies, 11*(2), 193–202. https://doi.org/10.1109/TLT.2017.2679009

Liu, S., Kang, L., Liu, Z., Fang, J., Yang, Z., Sun, J., Wang, M., & Hu, M. (2021). Computer-supported collaborative concept mapping: The impact of students' perceptions of collaboration on their knowledge understanding and behavioral patterns. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1927115

Liu, Y., & Chen, M. (2023). The knowledge structure and development trend in artificial intelligence based on latent feature topic model. *IEEE Transactions on Engineering Management*. https://doi.org/10.1109/TEM.2022.3232178

Liu, Z., Mu, R., Yang, Z., Peng, X., Liu, S., & Chen, J. (2022). Modeling temporal cognitive topic to uncover learners' concerns under different cognitive engagement patterns. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2022.2063904

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience, 34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

Mayordomo, R. M., & Onrubia, J. (2015). Work coordination and collaborative knowledge construction in a small group collaborative virtual task. *The Internet and Higher Education, 25*, 96–104. https://doi.org/10.1016/j.iheduc.2015.02.003

Niu, Y., Xie, R., Liu, Z., & Sun, M. (2017). Improved word representation learning with sememes. In R. Barzilay, & M. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 2049–2058). https://doi.org/10.18653/v1/P17-1187

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, ba and leadership: A unified model of dynamic knowledge creation. *Long Range Planning, 33*(1), 5–34. https://doi.org/10.1016/S0024-6301(99)00115-6

Novak, J. D., Bob Gowin, D., & Johansen, G. T. (1983). The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education, 67*(5), 625–645. https://doi.org/10.1002/sce.3730670511

Ouyang, F., Chen, Z., Cheng, M., Tang, Z., & Su, C.-Y. (2021). Exploring the effect of three scaffoldings on the collaborative problem-solving processes in China's higher education. *International Journal of Educational Technology in Higher Education, 18*(1), 35. https://doi.org/10.1186/s41239-021-00273-y

Ouyang, F., Wu, M., Zhang, L., Xu, W., Zheng, L., & Cukurova, M. (2023). Making strides towards AI-supported regulation of learning in collaborative knowledge construction. *Computers in Human Behavior, 142*, 1–11. https://doi.org/10.1016/j.chb.2023.107650

Patikorn, T., Deisadze, D., Grande, L., Yu, Z., & Heffernan, N. (2019). Generalizability of methods for imputing mathematical skills needed to solve problems from texts. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 396–405). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_33

Pazzani, M. J. (2000). Knowledge discovery from data? *IEEE Intelligent Systems and Their Applications, 15*(2), 10–12. https://doi.org/10.1109/5254.850821

Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers & Education, 143*, 1–14. https://doi.org/10.1016/j.compedu.2019.103673

Pfiffner, N. (2021). Identifying patterns in communication science: Mapping knowledge structures using semantic network analysis of keywords. In E. Segev (Ed.), *Semantic Network Analysis in Social Sciences* (pp. 192–215). Routledge. https://doi.org/10.4324/9781003120100-11

Phillips, M., Kovanović, V., Mitchell, I., & Gašević, D. (2019). The influence of discipline on teachers' knowledge and decision making. In B. Eagan, M. Misfeldt, & A. Siebert-Evenstone (Eds.), *Advances in Quantitative Ethnography* (pp. 177–188). Springer International Publishing. https://doi.org/10.1007/978-3-030-33232-7_15

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S., & Lee, D. (2021). Classifying math knowledge components via task-adaptive pre-trained BERT. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 408–419). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_33

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700

Sun, D., Ouyang, F., Li, Y., & Zhu, C. (2021). Comparing learners' knowledge, behaviors, and attitudes between two instructional modes of computer programming in secondary education. *International Journal of STEM Education, 8*(1), 1–15. https://doi.org/10.1186/s40594-021-00311-1

Sundararajan, N., Adesope, O., & Cavagnetto, A. (2018). The process of collaborative concept mapping in kindergarten and the effect on critical thinking skills. *Journal of STEM Education*, *19*(1). 5–13. Retrieved from https://www.learntechlib.org/p/182981/

Välimaa, J., & Hoffman, D. (2008). Knowledge society discourse and higher education. *Higher Education, 56*(3), 265–285. https://doi.org/10.1007/s10734-008-9123-7

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

de Weerd, J., Tan, E., & Stoyanov, S. (2017). Fostering interdisciplinary knowledge construction in computer-assisted collaborative concept mapping. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education* (pp. 391–396). Springer International Publishing. https://doi.org/10.1007/978-3-319-66610-5_32

Wen, Z., Gui, L., Wang, Q., Guo, M., Yu, X., Du, J., & Xu, R. (2022). Sememe knowledge and auxiliary information enhanced approach for sarcasm detection. *Information Processing & Management, 59*(3), 1–12. https://doi.org/10.1016/j.ipm.2022.102883

Wu, D., Yang, R., & Shen, C. (2021). Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm. *Journal of Intelligent Information Systems, 56*, 1–23. https://doi.org/10.1007/s10844-020-00597-7

Ye, Y., Qi, F., Liu, Z., & Sun, M. (2022). Going "deeper": Structured sememe prediction via transformer with tree attention. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 128–138). Retrieved from https://aclanthology.org/2022.findings-acl.12

Yeari, M., & van den Broek, P. (2016). A computational modeling of semantic knowledge in reading comprehension: Integrating the landscape model with latent semantic analysis. *Behavior Research Methods, 48*(3), 880–896. https://doi.org/10.3758/s13428-016-0749-6

Yoon, M., Lee, J., & Jo, I.-H. (2021). Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. *The Internet and Higher Education, 50*, 1–10. https://doi.org/10.1016/j.iheduc.2021.100806

Zabolotna, K., Malmberg, J., & Järvenoja, H. (2023). Examining the interplay of knowledge construction and group-level regulation in a computer-supported collaborative learning physics task. *Computers in Human Behavior, 138*, 1–17. https://doi.org/10.1016/j.chb.2022.107494

Zhang, S., Liu, Q., & Cai, Z. (2019). Exploring primary school teachers' technological pedagogical content knowledge (TPACK) in online collaborative discourse: An epistemic network analysis. *British Journal of Educational Technology, 50*(6), 3437–3455. https://doi.org/10.1111/bjet.12751

Zhao, J., Bai, T., Wei, Y., & Wu, B. (2022). PoetryBERT: Pre-training with sememe knowledge for classical Chinese poetry. In Y. Tan & Y. Shi (Eds.), *Data Mining and Big Data* (pp. 369–384). Springer Nature. https://doi.org/10.1007/978-981-19-8991-9_26

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ning Zhang Ph.D. student, College of Education, Zhejiang University. Her research interests are knowledge construction, learning analytics, online and blended learning.

FanOuyang research professor, College of Education, Zhejiang University. Her research interests are computer-supported collaborative learning, learninganalytics and educational data mining, AI in education, and online and blendedlearning.