DETECTING TOPICS OF CHAT DISCUSSIONS IN A COMPUTER SUPPORTED COLLABORATIVE LEARNING (CSCL) ENVIRONMENT

Dr. Gulgun AFACAN ADANIR
ORCID: https://orcid.org/0000-0002-0832-1808
Distance Education Center, Ankara University
Ankara, Turkey

ABSTRACT

Learning groups' conversations in computer supported collaborative learning (CSCL) environments result in significant information regarding the content of the course. This information is beneficial for instructors to analyze learners' activities during their collaboration process. In understanding these activities and performance of learners, the topic of conversation is important. The purpose of the study is to detect topics of chat discussions conducted by groups of learners while collaboratively studying in an online CSCL environment called Virtual Math Teams (VMT). We implemented the study in the context of a graduate level course during one term in a large state university in Turkey. Participants are MSc and PhD students registered to the course and divided over five groups of three students. We combined and employed methods of data mining, social network analysis, and topic detection to identify topics of learners' discussions. Our data analysis process aims to identify the task related topics occurred in chat discussion of learning teams. In our analysis we followed the stages of data preprocessing, segmentation analysis, and topic detection. Our purpose with the preprocessing stage was eliminating improper data for the main analysis and making the data ready for analysis stage. Therefore, our final corpus was shaped to involve 95% of initial chat messages. Segmentation analysis aims to explore organization of chat discussion and divides the chat logs into more manageable units according to their corresponding contents. In total, we resulted 294 segments including task related and non-task related ones. The topic detection analysis explored the content of chat segments and revealed the major subject of discussions with the use of latent semantic analysis, which is applied to find content similarity among segments and indicative words produced through the use of two mode network analysis.

INTRODUCTION

The important advantage provided by computer supported collaborative learning (CSCL) environments is that they automatically record user data (i.e. timestamps related to posting and receiving operations, name of sender, name of readers, content of messages) in log file format (Pozzi et al., 2007). The methods employed for the investigation of log files can be broadly categorized as quantitative or qualitative analysis methods, which are determined according to the type of the research (Bruckman, 2006). For instance, some researches investigate learner engagement in CSCL environments and considers quantitative indicators (Fournier, Kop, & Sitlia, 2011) such as learners' number of page accesses, time-on-task, resource use, as well as the number of the messages read, the postings made to a discussion board, and the frequency of the file up-loads (Retalis et al., 2006). The qualitative analysis of log files can be conducted with the use of appropriate frameworks such as grounded theory (Glaser & Strauss, 1971), and activity theory (Engestrom,

Miettinen, & Punamaki, 1999). In addition, analyzing the structure and the organization of collaborative processes are significant goals in the area of CSCL. In this regard, researchers offered various methods such as social network analysis, discourse analysis, and content analysis for the purpose of assessing collaborative process taking place in CSCL environments (Li, Wang, Liao, Zhao, & Huang, 2007).

Social network analysis (SNA) is defined as "the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities" (Mukherjee & Holder, 2004, p. 2). In general terms, SNA considers dynamics of the learner network and aims to analyze interactions occur in CSCL environments. That is, it mainly tries to reveal structure of the network by providing answers to questions: who collaborate with whom and who are the most active students during collaboration. SNA demonstrates the results of such questions by the use of graph theoretic constructs such as in-degree, out-degree, centrality, and eigenvalue metrics. In-degree indicates the number of chat utterances received by the student. Out-degree represents the number of chat utterances sent from one student to others during the collaboration. The degree of centrality is determined according to the relative significance of a vertex in the graph (e.g. the importance of an individual is in a social network) (Passmore, 2011). The eigenvalue metric is employed to reveal the influence of learners in the network. Some recent researches have applied SNA metrics to analyze the structure of collaborative interactions facilitated by CSCL environments. An Advanced System for Assessing Chat Participants (ASAP) is one of the initial studies employing SNA methods in CSCL. The ASAP study firstly conducted the preprocessing operation to chat data. Then, the study employed ranking process to calculate ranking of each learner according to number of interventions and score of utterances. It also applied SNA to reveal in-degree, out-degree, centrality, and eigenvalues metrics related to chat logs (Dascalu et al., 2008). Social networks adapting pedagogical practice (SNAPP) was offered as a monitoring/diagnostic tool that provides the functions of assessing the learners' interactions in discussion forums in learning management systems (LMS) (Bakharia, Heathcote, & Dawson, 2009). The SNAPP reveals the form of relationship among learners considering the links between their messages in a discussion board and provides a social network diagram, which demonstrates isolated students, patterns of instructor centered networks, group malfunctioning, and learners connecting small clustered networks and taking the role of information brokers. The LMSAnalytics tool was developed for investigating performance of individuals and groups in a networked learning environment (NLE) with the help of various approaches like descriptive statistics, social network analysis, content and context analysis. Regarding the social network analysis, the tool investigates interactions among peers, learner-tutor and learner-content in order to obtain a general overview of the collaborative learning activities mediated by the CSCL system (Petropoulou et al., 2010).

In the CSCL context, discourse analysis is generally employed to examine linguistic structures preferred by students for the coordination and organization of their interaction. Such methods consider theoretical frameworks such as speech-act theory (Searle, 1969) or dialogism (Bakhtin, 1986) with an effort to reveal communicative and coordinative functions of appropriate linguistic units within knowledge building discourse. One of the early automated techniques for coding dialogue acts was proposed by Erkens and Janssen (2008). Their method applied rule-based algorithm with consideration of keywords and sentence structures in order to detect communicative functions of chat messages, which are investigated under five categories which are argumentative, responsive, informative, elicitative, and imperative. Weinberger and Fischer (2006) proposed a multi-dimensional framework for the analysis of argumentative knowledge construction in CSCL environments. Their framework covers four different dimensions which are participation, epistemic, argument dimensions, and the dimension of social modes of co-construction. With the participation dimension, they aim to investigate learners' amount of contributions to collaborative work. The epistemic dimension investigates the discourse to understand whether the collaborative activities are task related or not. The argument dimension considers learners' construction and balance of arguments and counterarguments for the purpose of solving complex problems. The social modes of co-construction indicate that learners state contributions of their peers. Although Weinberger and Fischer (2006)

constructed a comprehensive framework for categorizing discourse elements, they did not transform it to an automated tool with the proper machine learning methods. In a recent study, Gweon, Raj and Rose (2011) aims to identify utterances by which learners perform "reasoning" in group discussions. As the initial step, the method detects the segments that consist of content in reasoning category which are examined under compare/contrast or cause/effect types. As the second step, every segment is categorized as one of five types such as theoretical concepts, prior knowledge, physical system properties, emergent system properties, and goals. Moreover, the study attempts to categorize the reasoning process into different groups. The study suggested that their classification can be employed by the help of machine learning methods.

The content analysis attempts to examine learner contributions in online discussion environments, and to detect the nature these contributions (Fournier et al., 2011). It employs categorization schemes to analyze linguistic content, and consist of the contents of graphical and other narrative resources. In their study, Retalis et al. (2006) developed a tool called CoSyLMSAnalytics to investigate learner activities in the Moodle LMS. Their method explores quantitative results related to user activities such as number of postings, number of replies, and types of posted messages. It also attempts to detect when teams talk about concepts or procedures to finalize the task, and whether teams complete the task collaboratively or cooperatively. Additionally, their approach conducts a path analysis by constructing clusters of learners who perform similar activities in one or more online sessions. Law et al. (2007) has offered a conceptual design for a learnable content and participation tool to be employed in the CSCL field. The toolkit mainly covers Preparatory, Analysis and Learning Mechanisms Components. Preparatory components transform discourse data to proper format required for the analysis, and present explanation of the coding schemes and coding rules. Analysis components provide statistics in the individual level, examine interpersonal interaction by providing measures of social network analysis and perform text analysis. Learning mechanism is dedicated for listing discourse segments and interpreting keywords, keywords concordance results, and results from the domain ontology analysis. Automatic Classification of Online Discussions with Extracted Attributes (ACODEA) framework was developed with the purpose of automatic analysis of online discussions (Mu, Stegmann, Mayfield, Rosé, & Fischer, 2012). The ACODEA framework is applied by the use of the SIDE tool. The framework is made of three major layers: (1) Extracting attributes, (2) Segmenting, (3) Coding. The first layer aims to identify textual features which are significant for achieving patterns related to machine learning algorithms. Rather than using formerly developed categories, the study has constructed a set of labels to classify collaborative tasks and activities. The Segmenting layer automatically divides the data into decided units by considering syntactic attributes, semantic attributes, and the unit of analysis. The Coding layer employs the multidimensional framework developed by Weinberger and Fischer (2006) categorizes the messages as claim, ground, warrant, inadequate claim, evaluation, prompts or empty message.

Topic detection methods attempt to discover the subject being talked in every thread of a chat conversation. In other words, the methods identify topics discussed in a specific time period, or examine the whole chat log and identify the topics that are discussed within it. Topic detection approaches consider supervised or unsupervised methods. In supervised methods, chat topics are firstly identified from training documents, then topics are assigned to documents in test category. In unsupervised methods, there is no earlier study to detect topics appear in documents. Alternatively, text documents are grouped according to their content similarity.

Supervised approaches have frequently employed Naïve Bayes, k-Nearest Neighbor, and Support Vector Machine (SVM) techniques in order to classify threads based on pre-determined topics. Ozyurt and Kose (2010) applied these three techniques together while identifying topics of Turkish chat messages. They proposed indicative feature sets for topics and classified messages in terms of these sets. In addition to the classification, they compared results of techniques and found the SVM having the best performance. Elnahrawy (2002)'s study categorized chat logs and newsgroup messages by employing Naïve Bayes, k-Nearest Neighbor, and Support Vector Machines (SVM) techniques. The results demonstrated that Naïve

Bayes classifier has the best performance among these three techniques. That is, the Naïve Bayes classifier requires less training time than the SVM technique, and requires less classification time than the k-Nearest Neighbor technique. Anjewierden, Kolloffel and Hulshof (2007) classified chat messages produced from the online collaborative work of learners. The study applied Naive Bayes classifiers for the purpose of identifying functional roles of messages as regulatory, domain, social, or technical types. Dong et al. (2006) identified topics of MSN messages by employing Naive Bayes, Associative Classification, and Support Vector Machine techniques. Similar to the study of Ozyurt and Kose (2010), the classification was performed based on formerly constructed indicative word sets which belong to topics. According to the results, SVM outperforms other two methods and generated better precision and accuracy values.

Unsupervised approaches utilized various techniques while classifying chat messages. For instance, Shen et al. (2006) aimed to detect threads in a message stream by considering three variations of the single-pass clustering algorithm and one new technique that additionally integrates linguistic features. Single-pass clustering algorithms focus on similarity of messages according to existing words and time distance among messages. Linguistic features cover sentence types and personal pronouns. According to the results, the proposed algorithm based on linguistic features is found the best one among the basic single-pass algorithm and its three variations. In their study, Wang and Oard (2009) offered to apply social and temporal contexts by the use of corresponding formulas. Social contexts are examined in two types: author and conversation context. The author context proposes that messages from the same person are most probably clustered into same conversation if they are temporally close to each other. The conversation context focuses on name mentions, and the temporal context considers the time distance while categorizing chat messages.

Wang et al. (2008), and Adams and Martell (2008) employed connectivity matrices to construct parent-child relationships between messages. Messages are initially defined with TF-IDF term vector representations for computing message similarity. In Wang et al. (2008)'s method links are established between two messages if their resemblance is beyond a threshold value. This value is calculated in terms of three different temporal features. Adams and Martell (2008) applied time-distance penalization, hypernym augmentation, and nickname augmentation for the purpose of detecting which message belongs to which thread. Results of the clustering demonstrated that time-distance penalization has the highest impact on increasing the performance of their algorithm. In their study, Mayfield, Adamson and Rosé (2012) applied a two-pass algorithm to detect three levels in a chat conversation: sentences, sequences and threads. In the initial pass, sentences are labeled in terms of features of unigrams, bigrams, and part-of-speech bigrams. Next, sequences are identified by utilizing a single-pass clustering algorithm. Every message is investigated according to the threshold value which is calculated by considering time distance and cosine similarity among messages. If the message is below the threshold, a new cluster is constructed. In the second pass, the sequences identified in the first pass are assigned to threads through cluster classifiers. Elsner and Charniak (2011) consider coherence models which examines text in its context. The entity grid model divides a document into entities and their syntactic roles, such as subject, object, other, and not-present. The role of each entity can be forecasted in terms of its former roles and number of occurrences. Topical entity grid model focuses on topic to word distributions besides earlier features. IBM-1 model aims to produce content words of the next sentence by considering the words of previous sentence. Pronouns and discourse newness are the other models. Additionally, time gap, speaker identity, and name mentioning features were employed. Trausan-Matu, Rebedea, Dragan and Alexandru (2007) proposed a tool for computer supported collaborative learning that can identify new topics when they are initiated during a conversation. Their approach considers detecting frequent words in the chat after irrelevant words are eliminated. Topics are detected by employing synonyms to detect common words, investigating patterns among the topics that are dynamically introduced, and based on user feedback. In their recent study, Trausan-Matu, Tascalu and Rebedea (2014) considered the "polyphonic framework" as a theoretical foundation and proposed the *PolyCAFe* system in order to analyze textual and gestural interactions occurring in collaborative groups. They

employed Natural Language Processing (NLP) for the identification of topics and connection between utterances.

Our review of the related work showed that prior studies perform the analysis of CSCL processes in various ways. SNA based studies typically consider the structure of the social network according to interactions of learners by employing some special metrics such as centrality, density, roles, groupings/cliques. Discourse analysis investigates chat messages in terms of their communicative and coordinative functions. Content analysis employs categorization patterns to investigate linguistic content of messages in addition to the contents of graphical and narrative resources. Topic detection analysis reveals the subjects being discussed in threads of chat discussions. Even though each of these methods contributes to offer significant perceptions to collaborative learning processes, to some extent they can reveal the structure and the organization of chat discussions in CSCL environments. In other words, these methods achieve the further importance if they are employed in a complementary way.

CSCL environments produce large amount of data related to learners' collaboration activities. Although such data provides significant input for the assessment of learning process, the analysis requires intensive effort and time allocation. Therefore, instructors and researchers need practical methods for the inspection of collaboration process since the assessment of only learners' final deliverable is not adequate for a complete and reliable evaluation of learners' collaboration. In this regard, our study aims to provide methods that consider chat discussions during collaboration process and reveal topics of discussions to be utilized by researchers and instructors for assessment.

In this study, we aimed to detect topics of chat conversations by employing methods of text mining, social network analysis and topic detection. Our study considered the chat conversations of learning groups while collaborating on assignments of a graduate level statistics course. Therefore, we focused on topics of speakers' conversations, which were shaped in the context of course assignments. For this purpose, we used supervised learning methods to investigate chat conversations in this study. In other words, in this analysis, we used one part of chat conversation as training set and identified the subjects, which are parallel to content of assignment questions. For the test set, we analyzed their topics considering subjects produced during the analysis of training set. We employed Latent Semantic Analysis (LSA) method to classify conversations. With this study, we provided methods for the analysis of learners' task related chat discussions. In other words, our method explores the portions of chat discussions in which student teams collaboratively worked to solve assignment items. That is, text mining methods match the discussion portions with each assignment item. By using the results of the methods, instructors and researchers can identify which item is discussed in which section of the chat discussion. Hence, they can consider the specific chat portion for investigation of learners' discussions based on particular assignment question. The remainder of the paper is organized as follows. In the section 2, we presented our methodology. We dedicated the section 3 for the results. In the final section, we presented the summary and implications of the results for researchers and practitioners.

METHODOLOGY

Research Design

The purpose of the study is to detect topics of chat discussions conducted by groups of learners while collaboratively studying in an online CSCL environment called Virtual Math Teams (VMT). Quantitative methods were applied in order to satisfy appropriate understanding for the research problem, which is stated as follows: "Which concepts are discussed within task-relevant chat segments of learning groups collaborated in VMT?"

Setting

We implemented the study in the context of a graduate level course during one term in a large state university in Turkey. The course covers basic concepts of empirical research and experimental design and structured in a way that the instruction was held face-to-face and assignments were collaboratively performed online. Totally, there were 15 registered students. Every student was assigned to a learning group randomly and five learning groups were constructed in total. Groups were obliged to perform course assignments by collaboratively studying online in the Virtual Math Teams (VMT) environment which was developed as a CSCL environment. The purpose of the activities was to enable students extend their understanding of major statistics concepts through collaborative assignments where they attempted to conduct a specific type of analysis by the use of SPSS software. In order to complete assignments, learning groups firstly performed online chat meetings, then published their findings as co-authored wiki documents.

During the term, learning groups collaborated on seven homework assignments and reported their works as a co-authored report in the online wiki environment. Students were required to collaborate with their group members by functions of the VMT system, therefore their entire communication could be investigated. The chat function of the VMT provided students with collaborating in a synchronous manner. In the chat environment, students could also utilize the whiteboard tool to clarify their solutions by drawing shapes or posting screenshots of their SPSS outputs. After the chat discussion, students were expected to summarize findings of their collaborative work as Wiki outputs which consist of textual information and graphical demonstrations.

Participants

Participants are MSc and PhD students of Informatics Institute in one of the state universities of Turkey. In total, fifteen students registered to the course and divided over five groups of three students. Since the course covers interdisciplinary subjects, it accepts students from diverse educational majors. The paper considers chat discussions of three teams: Team-1, Team-2, and Team-5. We selected these three teams since they carried out online collaboration sessions while working on assignments. The remaining teams demonstrated insufficient use of the chat tool, instead completed assignments by cooperatively and combining the individual works in VMT settings. Demographic characteristics of selected teams' students were provided in the Table 1.

Table 1. Demographic Characteristics of Students

Team-1			
Subject Handle	A_S	G_C	Y_A
Gender	Male	Female	Male
Grade	PhD	Masters	Masters
Graduate major	Biomedical Engineering	Cognitive Science	Cognitive Science
Team-2			
Subject Handle	H_A	Z_B	M_G
Gender	Male	Female	Male
Grade	PhD	PhD	PhD
Graduate major	Cognitive Science	Cognitive Science	Medical Informatics
Team-5			
Subject Handle	A_B	D_C	H_K
Gender	Male	Female	Male
Grade	PhD	Masters	PhD
Graduate major	Cognitive Science	Cognitive Science	Cognitive Science

Data Collection

We collected data related to learning process of each team after they completed whole assignments. These data consist of chat logs automatically generated by the VMT system which consists of textual message content as well as the teams' whiteboard activities together with author and time information. An excerpt from a sample log file together with its fields is depicted in the Table 2.

The line value demonstrates the order of the chat message and date indicates the date that the chat message is posted. There are three types of timing values. Start time shows the time that user begins to type the chat message whereas the time shows when that chat message is posted. The duration indicates the difference between start and post times of a chat message. Event type classifies learner activities as chat or whiteboard activity, or messages produced by the system. The last columns are allocated for the messages with the identification of users.

Table 2. Expert of a Log File

Line	Date	Start Time	Post Time	G_C	Y_A
25	11/13/2016	13:54	13:55	and gender is binary, whereas all IQ variables are ratio	
26	11/13/2016	13:54	13:55		gender is nominal
27	11/13/2016	13:55	13:55		brain volume, body height and body weight should be ratio variables as well

Data Analysis

The data analysis process aims to identify the task related topics occurred in chat discussion of learning teams. In our analysis we followed the stages of data preprocessing, segmentation analysis, and topic detection. Our purpose with the preprocessing stage was eliminating improper data for the main analysis and making the data ready for analysis stage. Segmentation analysis divides the chat logs into more manageable units according to their corresponding contents. The topic detection analysis investigates the content of chat segments and reveals the major subject of discussions.

We analyzed 71% of the entire data. That is, of the chat data, we considered the ones produced by teams 1, 2, and 5, which includes 6978 chat messages in total. The remaining data (i.e. data belong to team 3 and team 4) including 2735 chat messages were not analyzed in our study, mainly due to these teams' insufficient use of the chat tool to work on assignments. These teams instead completed assignments by cooperatively and combining the individual works in VMT settings.

Preprocessing Stage

The corpus had two main challenges for the analysis: it consists of non-English words and it has noisy structure because of misspellings and abbreviated use of words. Therefore, we conducted data preprocessing before starting to chat analysis. That is, we firstly removed non-English words from the corpus and then transformed misspelled words and abbreviations to their proper forms.

Segmentation Analysis

Segmentation analysis investigates how participants organize their chat interaction as long sequences (i.e. segments) involving a set of ordered chat messages. In this analysis, we inspected chat logs to detect activity

boundaries where new activities are initiated, hence existing activities are terminated or postponed (Zemel, Xhafa, & Cakir, 2007). The Figure 1 schematically represents the segmentation analysis.
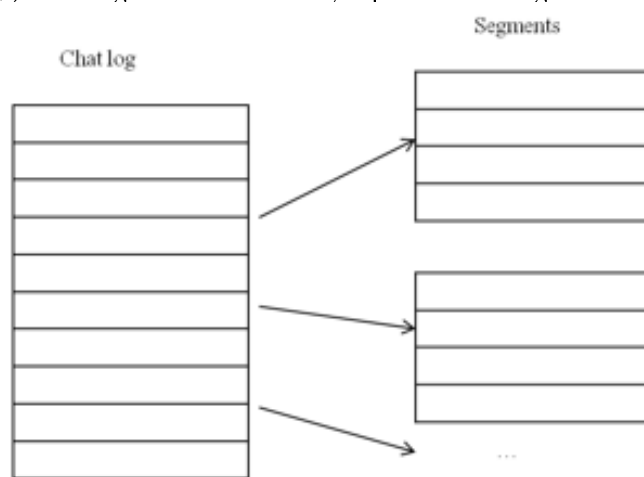

Figure 1. Representation of Segmentation Analysis

Considering the idea of segmentation analysis, we attempted to find segments within the conversation, hence focused on methods indicating the start of a new segment. In order to identify these methods or conventions, we examined chat conversations and explored frequent occurrences of words or phrases that usually appear at the beginning of segments or segment transition points (see Table 3). These methods comprise words and expressions that typically lead to transition between two succeeding segments. For example, the most frequent way of starting a new topic segment is one learner's offer to collaborate on a specific question. Since remaining learners in the chat room reply this proposal by posting responses (e.g. proposing solution possibilities), a conversation progressively starts to develop on the newly proposed topic.

Table 3. Segment Starting Methods

Method	Description	Example Chat Messages
1	'hi', 'hello', 'welcome' is a single word anywhere in the posting	'hello friends' 'hi there'
2	'let', 'let`s' is a single word anywhere in the posting	'let's deal with part g'
3	'part', 'for' is followed by question item/number in a posting	'in part b we should do partial correlation, am i wrong?' 'so, for 1d'
4	'question', 'q' is a single word anywhere in the posting	'question e, we need sphericity test and normality tests results'
5	'I think' is a phrase anywhere in the posting	'I think the best way is that we plot histogram and make analysis'
6	'how about', 'what about', 'by the way' is a phrase anywhere in the posting	'how about the goal of the study?' 'what about the variables?' 'by the way about 1b, do you have any ideas?'
7	'move', 'proceed', 'next', 'continue', 'pass' is a single word anywhere in the posting, sometimes prefixed with a "so"	'so, we can move on to the outlier one' 'If you agree, we can pass to the second question'
8	'start', 'finish', 'stop' is a single word anywhere in the posting, typically prefixed with "let's"	'let's start from 3rd question ok?' 'let's stop here'
9	question item or number anywhere in the posting	'ok, f'

Topic Detection

The segmentation analysis brings us to identify topics in chat discussions of learning groups. Every segment points to a specific topic, which is usually related to the question of the assignment the group is performing collaboration. Chat topics are broadly classified as task related and non-task related. Task related topics indicate discussions that cover solutions provided to questions of assignments. One example task related discussion is that members of a learning group try to interpret normality test results provided by the statistics tool. The topics are treated as non-task related if the group's messages are related to the socializing or sharing experiences about daily lives. For instance, one group initiated the conversation by salutations, then talked about their overloaded duties such as exams or other assignments. After this talk, they initiated to collaborate on course work and talked about social life during breaks. Chat conversations may flow on various task related topics, as well as non-task related topics may exist between task related topics. The organization of topics develops in terms of social dynamics of the learning group.

In our data, topics of task related segments were formed in parallel to themes extracted from questions of assignments. In order to detect topic of each segment, we focused on keywords that the teams utilized while solving questions. For our purpose, we firstly applied two mode network analysis to explore terms that teams employed while collaboratively solving each question of each assignment. The goal of this network is to represent the relation between two different nodes such as teams and terms. Nodes indicating teams are placed at the center and nodes representing terms are placed around the network. A line is constructed between a team node and a term node if the term is utilized by the team. The weight of the line shows the number of occurrences of the term in the discussion of the team. We converted each two mode network to its resultant one mode network to identify mutual words employed by teams while collaborating on a solution of the question. For example, for the question-c of assignment-1, mutual words were detected as condition, dependent, independent, time, and variable, which are as a result identified as indicative terms of this question. In the same manner, we identified indicative words for every question of seven assignments.
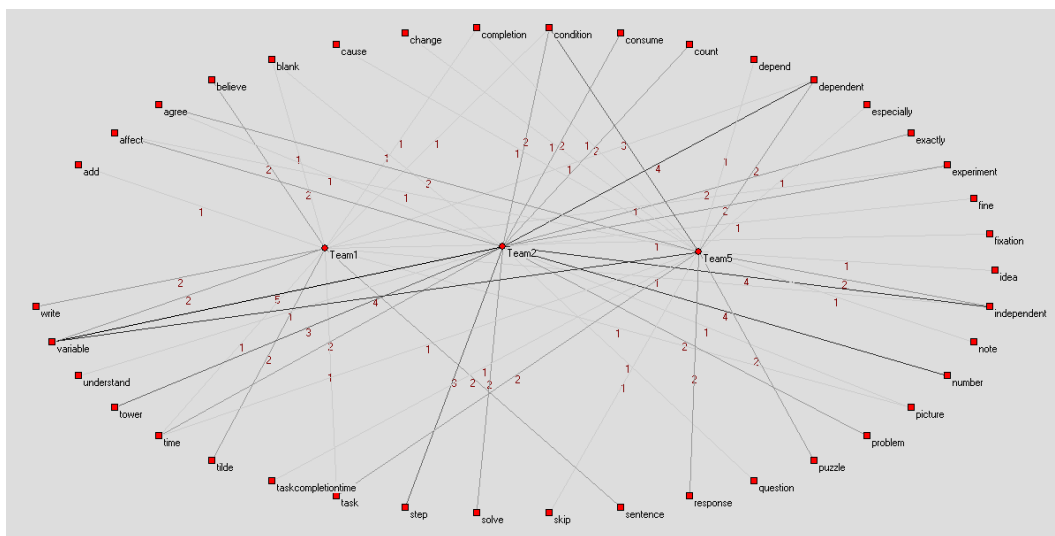

Figure 2. Two Mode Network that Shows Terms Used by Teams While Solving
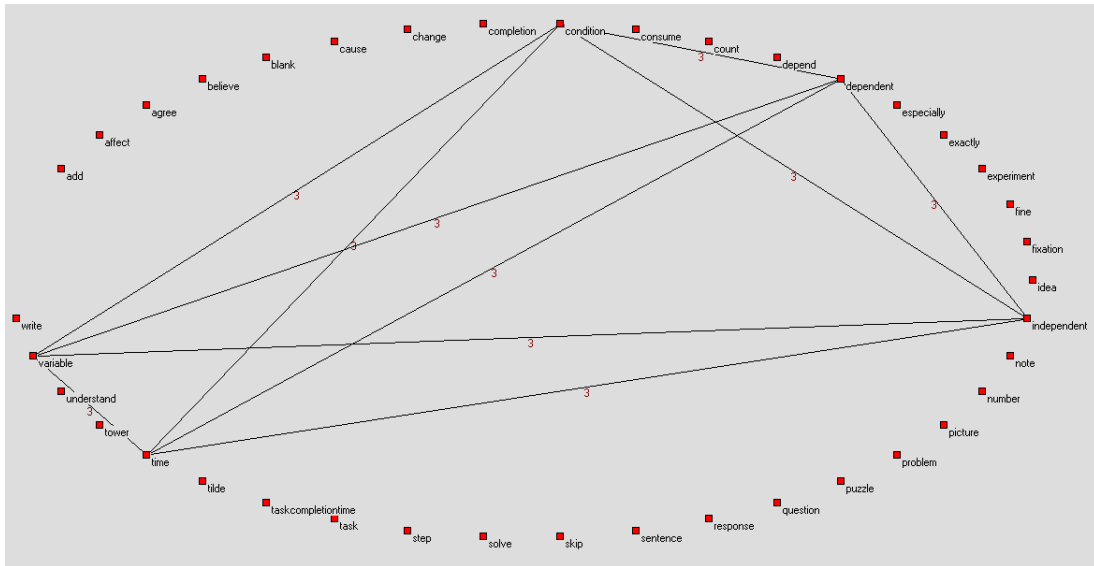Assignment-1 Question-c

Figure 3. One Mode Network that Shows Mutual Words Utilized for the Solution of the Assignment-1 Question-c

Our next step is to identify which segment maps to which question, we compared text of each segment (Sn) with indicative terms of each question (Qn). More specifically, we applied the latent semantic analysis (LSA) method to reveal most probable mappings among segments and questions in terms of their semantic similarity. LSA is defined as "a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse." (Landauer, Foltz, & Laham, 1998, p. 8). In order to apply LSA, we followed two major steps.

In the initial step of LSA, we constructed a term-document matrix that its rows demonstrate vectors of all terms appeared in segments and its columns demonstrate the vectors of all questions and segments. The term-document matrix is represented by a matrix A, with m x n dimensions, pointing to terms, and questions and segments respectively. Every cell value can be represented as A[i, j]=a that the ith term occurs in the jth question or segment for a times. For instance, we generated the following matrix in Figure 4 for the chat segments of team-5 and the question documents in the context of assignment-1.

	Documents										
Terms	Q1a	Q1b	Q1c	Q1c2	Q1d	Q2a	s3	s4	s5	s6	s10
answer	0	0	0	0	0	0	0	0	0	0	1
categorical	0	0	0	0	0	0	0	0	0	1	0
chart	0	0	0	0	0	1	0	0	0	0	0
collect	0	0	0	0	0	0	0	1	0	0	0
condition	0	0	0	0	0	0	0	1	1	0	2
data	0	0	0	0	0	1	0	1	0	0	3
dependent	0	0	0	1	0	0	0	0	1	0	1
descriptive	0	0	0	0	0	1	0	0	0	0	0
design	0	0	0	0	0	0	0	1	0	0	0
experiment	0	1	0	0	0	0	0	2	0	0	0
…											

Figure 4. Chat Segments of Team-5 and the Question Documents in the Context of Assignment-1

In the second step of LSA, we employed singular value decomposition (SVD) and acquired finalized vector forms of questions and segments. SVD considers the following formula for the matrix A:

$$A = S\Sigma U^T$$

Where the S matrix involves the eigenvectors of matrix B, the U matrix involves eigenvectors of C, and the Σ matrix involves the singular values obtained as square roots of the eigenvalues of the B matrix. Singular values are ignored and replaced with 0 when they are too small. In other words, k singular values are kept in Σ, which is reduced to a Σ_k matrix. Similar to this transformation, S and U^T are reduced to S_k and U_k^T matrices correspondingly, and the matrix A is approximated as follows:

$$A_k = S_k \Sigma_k U_k^T$$

After this transformation, documents and terms acquire new representations. Documents are represented with column vectors - $\Sigma_k U_k^T$ and terms are represented with row vectors - $S_k \Sigma$.

In the final step of topic detection, we calculated cosine similarity between their finalized vectors to reveal semantic similarity among segments and questions. Cosine values range from -1 to 1. The values close to 1 indicate high degree of semantic similarity. In this way, our study revealed which segment maps to which question.

RESULTS

Preprocessing Results

As the initial step of preprocessing, we grouped chat messages according to their language content. Although 95 % of the discussions were held in English, learners employed Turkish for social topics, typically including non-task related messages. This was an expected manner since students may prefer to use their native language for socializing. Since we focus on conversations on task related issues, we eliminated chat data in Turkish language and kept the ones in English language. At the end, the final corpus was formed to cover 95% of the chat messages produced by teams 1, 2, and 5. The noisiness in data was due to misspellings and abbreviations. Hence, as the second step, we transformed misspelled words and abbreviations to their proper forms. When we completed these procedures, the data became ready for successive steps of our analysis.

Segmentation Analysis Results

By applying segment starting methods, we detected segments in chat logs of assignments. Teams have conducted numerous online chat meetings, therefore various segments were generated in these sessions. Teams' number of segments distributed over assignments was provided in Table 4.

Table 4. Distribution of Segments over Assignments

	Assignment1	Assignment2	Assignment3	Assignment4	Assignment5	Assignment6	Assignment7
Team-1	10	24	8	16	8	28	8
Team-2	15	5	15	5	5		9
Team-5	54	10	27	23	9	11	4

These segments consist of learners' discussions to accomplish requirements of the assignment as well as coordination related issues. Among these segments, we consider the ones related to the learners' studies in solving the assignment questions in order to explore teams' studies on major concepts of the course. The teams' total number of task related and non-task related segments are provided in the Table 5. According to these results, we analyzed the task related segments, which is approximately 54% of whole segments.

Table 5. Distribution of Segments

Team	Number of task related segments	Number of non-task related segments
1	48	54
2	37	17
5	74	64
Total	159 (%54)	135 (%46)

For the investigation of reliability of segmentation analysis, we employed two coders (first author and one researcher) and computed two different indices proposed in the research of Strijbos and Stahl (2007):

➢ For the assignment of a thread or not by both coders (% thread);
➢ For the assignment of the same thread whenever both assigned a thread (% same).

Table 6 presents results of the reliability trial consisting of 159 chat lines for the pair of coders.

Table 6. Results of the Reliability Trial

Pair	% Thread	% Same
1-2	.72	.77

The CSCL area doesn't propose a threshold value as the agreement reliability of segmentation (De Wever et al., 2006; Rourke, Anderson, Garrison, & Archer, 2001), nor the area of content analysis (Riffe, Lacy, & Fico, 1998). Considering various views, a range of .70 -.80 for proportion agreement can be employed as the criterion value. The results showed that both coders identified a thread in 72 % of all cases and 77 % of assignments are identical. Based on these results, the reliability of segmentation analysis was found to fit the range of .70 -.80.

Topic Detection Results

For the purpose of detecting topic of each task related segment, we focused on keywords that teams utilized while collaborating on the relevant question of the assignment. As explained in the methodology, we employed two mode network analysis to explore terms which teams employed while collaboratively solving each question of each assignment. Among these words, we considered the mutual ones utilized by three teams by developing one mode networks from two mode networks. In the final, one mode networks provided us with the keywords of questions, which are demonstrated in Table 7. For instance, we revealed the keywords of question 1a of assignment-1 as minimum, number, and question. These results revealed that these three words are the important and commonly used related to this question, hence provides clues for the match of corresponding segments with this question.

Table 7. Keywords Detected for Questions of Assignments

Assignment-1	Keywords
1a	Minimum, number, question
1b	Experiment, participant, study, variable
1c	Goal, problem, time, condition, dependent, independent
1d	Nominal, ratio, variable
2a	Chart, data, descriptive, frequency, graph, split, value
2b	Descriptive, histogram, plot, toh
2c	Condition, data, distribute, group, mean, normally, normality, picture, puzzle, result, sigma, significant, split, toh, value
2d	Outlier, detection
2e	Log, transformation

Assignment-2	Keywords
1	Brain, fit, heavy, interval, ratio, variable, volume, weather, winter
2	Normality, table, test
3a	Bivariate, correlation, mean, mrivolume, partial, pearson, square, sum, weight
3b	Coefficient, correlation, height, positive, time, total, weight
4b	Fit, graph, significance, significant
4d	Residual
4e	Fit, line, part
4f	Model, mrivolume, predictor, value

Assignment-3	Keywords
a	age, baseline, fit, mean, predict, significance, statistic, table, value, variable
b	equation, model, number, predict
c	answer, classification, table, wrong
d	category, constant, contribution, equal, odds, outcome, predict, ratio, significance, statusquo, variable, wald
e	affect, age, odds, predictor, probability
f	association, confidence, interpret, interval, model, odds, positive, predict, probability
g	cooks, dfbeta, distribute
h	independent, multicollinearity

Assignment-4	Keywords
1	dependent, independent, measure, post, time
2	analysis, dependent, distribution, nonsignificant, normal, normality, posttest, pretest, theorem, variance, variable, homogeneity, levene, post, test
3	anova , difference, group, hypothesis, reject, significance
4	anova, difference, enter, mean, significance, table, test
5	difference, group, mean, normality, paired, posttest, pretest, ttest

Assignment-5	Keywords
a	dependent, independent, mood, stoprule, variable
b	factorial, independent, measure, mood
c	anova, assumption, data, distribute, kolmogorov, listcount, normality, normally, parametric, positive, result, separately, test
d	box, effect, mean, step
e	hoc, post, result

Assignment-6	Keywords
a	data, understand
c	condition, dependent, factor, independent, measure, variable
d	condition, effect, experiment, order, participant
e	assumption, difference, graph, mauchlys, normal, normality, parametric, significance, substantial, variance
f	condition, grand, mean, score, separate, ssb, sst
g	comparison, effect, eta, partial
h	anova, friedman, significant

Assignment-7	Keywords
a	dependent, experiment, independent, list, score, variable
b	assumption, box, dependent, equal, homogeneity, levene, manova, normality, sample, significant, test, time, variable, dependent, explore, list, normal, normality, year
c	covariance, look, multivariate, pillai, test
d	anova, bonferroni, contrast, dependent, difference, error, group, hoc, manova, post, posthoc, significant, sphericity, tukey, variance
e	range, variate, year

In order to explore which segment maps to which question, we compared each segment (sn) with keywords of each question (qn). For this comparison, we applied the latent semantic analysis (lsa) to compute the semantic similarity between segments and keywords. The LSA results demonstrate the cosine similarity between segments and questions. The value of cosine similarity changes from 0 to 1. The values near to 1 show a higher level of similarity between the segment and the question. In this way, we can explore which chat segment relates to which question statement (i.e. topic). The Table 8 demonstrates the match that we found among segments and questions of assignments related to collaborative activities of Team-1.

Table 8. Team-1's Segments and Matching Questions

Assignment1											
Segment	s2	s4	s5	s6	s10						
Question	1a	1b	1c	1d	2a						
Assignment2											
Segment	s3	s7	s9	s14	s16	s17	s21	s23	s24		
Question	1	2	3a	3a	3b	3b	4b	4d	4e		
Assignment3											
Segment	s2	s4	s5	s7	s8						
Question	b			a	b						
Assignment4											
Segment	s3	s6	s8	s9	s14	s16					
Question	1	2	2	2	3	4					
Assignment5											
Segment	s2	s3	s5	s6	s8						
Question	a		b	c	c						
Assignment6											
Segment	s2	s4	s6	s7	s9	s11	s18	s20	s23	s25	s27	s28
Question	d	d	d	e	d	e	g	g	g	f	g	h
Assignment7											
Segment	s2	s3	s4	s6	s7	s8					
Question	a	b	b	b	c	c					

When we analyzed the topics of questions we identified that related to the first assignment, Team-1 discussed about "Steps of the experiment" in s2, "Design of the study" in s4, "Variables and goal of the study" in s5, "Scales of variables" in s6, and "Descriptives" in s10. Related to the second assignment, Team-1 discussed about "Design of the study" in s3, "Descriptives and Test of Normality" in s7, "Correlation" in s9, s14, s16, s17, "Model Fit" in s21, "Residual" in s23, and "Regression" in s24. Related to the third assignment, Team-1 discussed about "Model Equation" in s2 and s8, "Model fit" in s7. Related to the fourth assignment, Team-1 discussed about "Variables" in s3, "Assumptions" in s6, s8, s9, and "Statistical test" in s16. Related

to the fifth assignment, Team-1 discussed about "Variables of the Study" in s2, "Design of the study" in s5, and "Assumptions of Anova" in s6 and s8. Related to the sixth assignment, Team-1 discussed about "Counterbalancing" in s2, s4, s6, and s9, "Assumptions of Anova" in s7, s11, PostHoc Test in s18, s20, s23, s27, "Applying Anova" in s25, and "Nonparametric Test" in s28Related to the seventh assignment, Team-1 discussed about "Variables of the Study" in s2, "Assumptions of Manova" in s3, s4, s6, and "Multivariate Tests" in s7, s8.

DISCUSSION AND CONCLUSION

Important advantage of CSCL environments is that they generate system logs that keep records of interactions occurring among learners. Since these logs detect instances where students provide questions, seek information and conduct reasoning together, learning activities become visible to the instructors. The increasing use of computer-mediated communication tools such as social networking, chat, instant messengers and wikis as elements of CSCL applications led to big repositories of such learning interactions. Yet, analyzing hundreds of lines of collaborative interactions of learning teams becomes a time consuming task for researchers and instructors. On the other hand, a thorough monitoring of the collaboration process is essential to support teachers to implement a fair assessment of group work and provide support when needed (Wang, 2009).

In this study, we attempted to bring ideas from data mining and social network analysis methods that will explore discussion topics in parallel to content of the course. For this purpose, we initially employed linguistic markers in the beginning of a discussion in chat to indicate segment boundaries. This step satisfied the necessary pre-processing to enhance the document similarity analysis performed in the next phase. The keywords used in question statements can be applied at this moment to navigate through chat logs. The segments consistent with questions provide teachers with detecting those interactional episodes where learning groups collaborated on key statistical concepts.

Chat logs were generated for each assignment individually and they cover various topics in mixed format. Therefore, one cannot directly recognize which question was discussed in which portion of the discussion. Hence, we aimed to split chat logs in terms of their focus and detect the topic of each chat part. For the division of chat logs, we used a set of segment starting methods that aimed to initiate new topics in chat. By employing these methods, we could detect 90% of segment starting messages. However, remaining messages including these keywords were detected as false segment beginners. Therefore, after we marked keywords in the whole chat corpus, we manually checked marked messages for revealing exact segment boundaries. Similar to our segmentation approach, Khan, Fisher, Shuler, Wu, & Pottenger (2002) employed patterns to explore start of conversation threads. These patterns were produced after observations of human experts and classified as positive and negative patterns. By the help of positive patterns, they positively realize starts of threads. Their positive patterns are as follows: Question mark in the posting, "Who, what, when, where, how, why" words in the posting, "Do Did Does Are Is Were Was Shall Will Can Could Would Have Had Has" words in the posting, a NAME in the posting, One of "hi, hey, yo, hello" is the beginning of a posting. Yet, applying only positive patterns result in some false positives which don't actually specify the beginning of a thread. This study was consistent with our results on deceptive segment starters. We eliminated the false positives after manually checking each starting message.

For identifying topics of segments, we considered frequent keywords that learning groups employed while collaborating on solving questions. These keywords are expected to capture the main idea of the questions. By applying two mode network analysis, we identified lists of indicative words for each question. In parallel to our methodology, Ozyurt and Kose (2010) have explored indicative words for various topics of chat conversations and utilized supervised methods to classify chat topics. Their methodology covered feature selection for the identification of indicative words and terms. Since their study covered general chat

discussions, the suggested indicative words belong to topics like sports, flirting, entertainment, etc. With the selected methods, they can identify topics of 154 conversations as sport, love, entertainment, education, or slang and concluded 87% accuracy rate in average. On the other hand, we implemented our study in the context of a course, so our keywords are consistent with questions of assignments. More importantly, our approach considering two mode network analysis to explore indicative words of assignment questions is a novel approach for the CSCL and topic detection area. In order to find mappings among segments and questions (i.e. topic), we performed a comparison by applying latent semantic analysis (LSA). Thus, we identified chat segments with their related questions and concepts as the major idea of questions. In terms of literature review, supervised methods usually preferred to apply Naïve Bayes, k-NN, and SVM algorithms while categorizing chat topics. In this respect, our methodology is different from the existing studies.

As the future study, collaborative tool designers can propose a teacher dashboard that involves modules in consistent with our methodology. In the first module, the instructor uploads the log file covering chat messages of learning groups. Then, the system marks potential segment starting messages by the help of keywords and phrases offered as the segment starting methods. The instructor should check whole starting messages and confirm the actual ones. Based on confirmed messages, the system conducts segmentation analysis, hence chat segments can be generated. In order to compare segments with questions, the module produces indicative words of questions and employs LSA. Additionally, the system provides the list of topics discussed by teams while collaborating on questions of the assignment. In this way, the instructor can choose topics to demonstrate and review the matching chat log.

BIODATA and CONTACT ADDRESSES of AUTHOR

Dr. Gulgun AFACAN ADANIR is instructor of Ankara University Distance Education Center. She also supports instructional activities of Ankara University Informatics Department. She gained her Ph.D from Middle East Technical University (METU) Information Systems at February, 2016. She holds a BSc in Computer Education and Instructional Technology (2006), an MSc in Information Systems (2009) both from Middle East Technical University. Her research interests are in distance education, learning analytics, e-government, and instructional technology.

Gulgun AFACAN ADANIR
Ankara University, Distance Education Center
Address: Ankara University Distance Education Center, 06830, Golbasi, Ankara, Turkey
Phone: +90 3126000141,
E-mail: gafacan@ankara.edu.tr

REFERENCES

Adams, P. H., & Martell, C. H. (2008, August). Topic detection and extraction in chat. In *Semantic Computing, 2008 IEEE International Conference on* (pp. 581-588). IEEE.

Anjewierden, A., Kolloffel, B., & Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *International Workshop on Applying Data Mining in e-Learning (ADML 2007)*.

Bakharia, A., Heathcote, E., & Dawson, S. (2009). *Social networks adapting pedagogical practice: SNAPP* (Doctoral dissertation, University of Auckland, Auckland University of Technology, and Australasian Society for Computers in Learning in Tertiary Education (ascilite)).

Bakhtin, M. (1986). The problem of speech genres (V. McGee, Trans.). In C. Emerson & M. Holquist (Eds.), *Speech genres and other late essays* (pp. 60-102). Austin: Univ. of Texas Press.

Bruckman, A. (2006). Analysis of log file data to understand behavior and learning in an online community. In *The International handbook of virtual learning environments* (pp. 1449-1465). Springer Netherlands.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & education, 46*(1), 6-28.

Dascalu, M., Chioasca, E. V., & Trausan-Matu, S. (2008). ASAP-An Advanced System for Assessing Chat Participants. In *Artificial Intelligence: Methodology, Systems, and Applications* (pp. 58-68). Springer Berlin Heidelberg.

Dong, H., Cheung Hui, S., & He, Y. (2006). Structural analysis of chat messages for topic detection. *Online Information Review, 30*(5), 496-516.

Elnahrawy, E. (2002, November). Log-based chat room monitoring using text categorization: A comparative study. In *The International Conference on Information and Knowledge Sharing, US Virgin Islands*.

Elsner, M., & Charniak, E. (2011, June). Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 1179-1189). Association for Computational Linguistics.

Engestrom, Y., Miettinen, R., & Punamaki, R. L. (1999). *Perspectives on activity theory*. Cambridge University Press.

Erkens, G., & Janssen, J. (2008). Automatic coding of dialogue acts in collaboration protocols. *International journal of computer-supported collaborative learning, 3*(4), 447-470.

Fournier, H., Kop, R., & Sitlia, H. (2011). The value of learning analytics to networked learning on a Personal Learning Environment.

Glaser, B. S., & Strauss, A. (1971). A. 1967, The discovery of grounded theory. *New york*.

Gweon, Raj, B., & Rosé, C. P. (2011). The automatic assessment of knowledge integration processes in project teams. In *Proceedings of Computer Supported Collaborative Learning* (pp. 462-469).

Janssen, J., Erkensa, G., Kanselaara, G., & Jaspersa, J. (2007). Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Computers & Education, 49*(4), 1037–1065.

Khan, F. M., Fisher, T. A., Shuler, L., Wu, T., & Pottenger, W. M. (2002). Mining chat-room conversations for social and semantic interactions. *Computer Science and Engineering, Lehigh University*.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25*(2-3), 259-284.

Law, N., Yuen, J., Huang, R., Li, Y., & Pan, N. (2007, July). A learnable content & participation analysis toolkit for assessing CSCL learning outcomes and processes. In *Proceedings of the 8th international conference on Computer supported collaborative learning* (pp. 411-420). International Society of the Learning Sciences.

Li, Y., Wang, J., Liao, J., Zhao, D., & Huang, R. (2007, July).Assessing collaborative process in cscl with an intelligent content analysis toolkit. In *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on* (pp. 257-261). IEEE.

Mayfield, E., Adamson, D., & Rosé, C. P. (2012, July). Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 60-69). Association for Computational Linguistics.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning, 7*(2), 285-305.

Mukherjee, M., & Holder, L. B. (2004, August). Graph-based data mining on social networks. In *Workshop on Link Analysis and Group Detection (LinkKDD2004)*.

Ozyurt, O., & Kose, C. (2010). Chat mining: Automatically determination of chat conversations' topic in Turkish text based chat mediums. *Expert Systems with Applications, 37*(12), 8705-8710.

Passmore, D. L. (2011). Social network analysis: Theory and applications.*Tersedia: http://code. pediapress. com/[12 Juni 2014].*

Petropoulou, O., Altanis, I., Retalis, S., Nicolaou, C. A., Kannas, C., Vasiliadou, M., & Pattis, I. (2010). Building a tool to help teachers analyse learners' interactions in a networked learning environment. *Educational Media International, 47*(3), 231-246.

Pozzi, F., Manca, S., Persico, D., & Sarti, L. (2007). A general framework for tracking and analysing learning processes in computer- supported collaborative learning environments. *Innovations in Education and Teaching International, 44*(2), 169-179.

Retalis, S., Papasalouros, A., Psaromiligkos, Y., Siscos, S., & Kargidis, T. (2006). Towards Networked Learning Analytics–A concept and a tool. In *Proceedings of the fifth international conference on networked learning.*

Riffe, D., Lacy, S., & Fico, F. (1998). Analyzing media messages: Quantitative content analysis. *Mahwah. NJ (USA): Lawrence Erlbaum Associates.*

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International journal of artificial intelligence in education (IJAIED), 12*, 8-22.

Searle, J. R. (1969). Speech Act Theory. Cambridge, MA: Cambridge University Press.

Shen, D., Yang, Q., Sun, J. T., & Chen, Z. (2006, August). Thread detection in dynamic text message streams. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*(pp. 35-42). ACM.

Strijbos, J. W., & Stahl, G. (2007). Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication.*Learning and Instruction, 17*(4), 394-404.

Trausan-Matu, S., Rebedea, T., Dragan, A., & Alexandru, C. (2007). Visualisation of learners' contributions in chat conversations. *Blended learning*, 217-226.

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2014). PolyCAFe—automatic support for the polyphonic analysis of CSCL chats. *International Journal of Computer-Supported Collaborative Learning, 9*(2), 127-156.

Uthus, D. C., & Aha, D. W. (2013). Multiparticipant chat analysis: A survey. *Artificial Intelligence, 199*, 106-121.

Wang, L., & Oard, D. W. (2009, May). Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics* (pp. 200-208). Association for Computational Linguistics.

Wang, Y. C., Joshi, M., Cohen, W. W., & Rosé, C. P. (2008, March). Recovering Implicit Thread Structure in Newsgroup Style Conversations. In *ICWSM*.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71-95.

Zemel, A., Xhafa, F., & Cakir, M. (2007). What's in the mix? Combining coding and conversation analysis to investigate chat-based problem solving. *Learning and Instruction, 17*(4), 405-415.