Analyzing Peer Interactions in Computer-Supported Collaborative Learning: Model, Method and Tool*

Yanyan Li, and Ronghuai Huang

Knowledge Science & Engineering Institute, School of Education Technology,
Beijing Normal University, 100875, Beijing, China
Email: liyy1114@gmail.com

Abstract. One of the most important facets in CSCL research is the interaction between individual and collaborative learning activities. This paper proposes a holistic and complementary analysis model of the collaborative interactions base on three dimensions – process pattern, social relationship and topic space. By making use of content analysis, social network analysis and text mining technologies, asynchronous discussion transcripts are semi-automatically processed to address the questions concerned with peer interactions in collaborative learning: what are they talking about, who are talking to whom, and how do they talking with others. An integrated tool with comprehensive functionalities is designed and implemented to support collaborative interaction analysis with intelligence and visualization features. With the assistance of the tool, a case study is conducted to analyze the discussion records of a class composed of 18 graduate students who enrolled in a course along with online discussion in knowledge forum platform.

Keywords: CSCL, Interaction Analysis, Text Mining, Content Analysis Tool, Social Network Analysis

1 Introduction

Currently, there is a growing adoption of computer-based facilities in educational practice to foster online collaboration. This practice is commonly described as the field of Computer Supported Collaborative Learning (CSCL). In CSCL environments, online asynchronous discussion takes a central place, which allows learners to share information, exchange ideas, address problems and discuss on specific themes. All exchanges of information between students are stored in the discussion transcripts. These transcripts can be used by teachers and students for reflection purposes or they can serve as data for research [9]. This asynchronous interaction, confined in the transcripts of the discussion, is thus the object of a large body of recent educational research.

* The research work was supported by the National Science Foundation of China (NSFC: 60705023)

So far, several approaches have been put forth to analyze interactions in the computer-supported collaborative learning (CSCL). The typical methods include analysis of computer-generated quantitative log files, social network analysis, discourse analysis and content analysis [4]. Quantitative log files generated and stored in the CSCL environments serve as an easily accessible source for analyzing collaborative process, but Nurmela et al. point out that researchers should not heavily depend on the information recorded in log files, but to combine this with an analysis of its content, especially the content of collaborative dialog or discourse [11]. Social network analysis (SNA) is usually used to study the way people participated and interacted with each other, especially investigates the relationship between participants rather than the discussion content. Discourse analysis is a broad and complex interdisciplinary field involving linguistics, anthropology, and sociology, which focuses on studying the naturally-occurring speech or conversation in context. But the indetermination of context causes the difficulty for understanding language use. Content analysis is defined as "a research methodology that builds on procedures to make valid inferences from text" [2]. Wever et al. [16] give an overview of different content analysis schemes that reflect the diversity in the theoretical base, the amount of information about validity and reliability, and the choice for the unit of analysis. Compared with other methods, content analysis is widely used to analyze and assess the collaborative interaction. The traditional content analysis mainly depends on the manual coding, which is time-consuming and tedious for the researchers. So it is indispensable to make use of tools to facilitate coding process for the interaction discourse analysis.

The aim of this kind of research is to provide a more complete picture of peer interaction in CSCL based on interaction analysis. We believe that these understandings will contribute to the development of better pedagogical frameworks and software that more effectively support learning and tutoring by design. Therefore, by incorporating content analysis, text mining, and social network analysis, this paper proposes a multidimensional analysis model to study peer interactions, expecting to provide an integrated foundation for in-depth investigation of collaborative learning in CSCL. The three methods are used to triangulate and contextualize our findings and to stay close or connected to the first-hand experiences of the participants themselves. Furthermore, this paper describes the design and implementation of an intelligent content analysis tool by adopting the quantitative statistics for participation and interaction analysis, and text analysis in addition to the semi-automatic coding support.

2 A Multidimensional Analysis Model to Study Collaborative Interaction

Discussion boards are one of the most commonly used facilities to support collaborative learning. Asynchronous text-based discussions present several advantages as compared to synchronous discussions: students get more opportunities to interact with each other and students have more time to reflect, think, and search for extra information before contributing to the discussion [17]. The facts that all

communication elements are made explicit in the written contributions to the discussions. By browsing the discussion transcripts, teachers and researchers are mostly interested in the following three questions.

- How do the students talking with others?
- What are the students talking about?
- Who are talking to whom?

When the teachers find out the answers to the questions, it is helpful for them to further understand the students' collaborative process, discover the possible existing problems with regard to the collaborative interaction, and accordingly take the necessary intervention strategies to facilitate collaborative learning.

Inspired by the three questions, we propose a multidimensional research model for studying peer interaction in CSCL, as shown in figure 1. The model comprises three dimensions: Process Pattern, Topic Space, and Social Network. Process Pattern reflects the interaction patterns among students involved in the collaborative learning, which emphasizes probing the diverse speech intentions and changing trend. Topic space reflects the knowledge or concepts that students used in collaborative learning process, which focuses on recognizing the discussion topics emerged in the participants' interaction. Social relationship reflects the dynamic mechanism influencing knowledge flow. It is to find out the relationship between the participants in collaborative learning, and the diverse roles the participants play to fulfill a specified task.

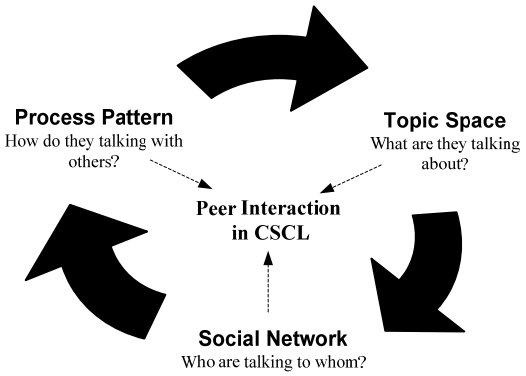

Fig. 1. A Multidimensional research model for studying collaborative interaction.

3 Methods

3.1 Content Analysis

Content analysis is often adopted to unlock the information captured in transcripts of asynchronous discussion groups with the aim to reveal information that is not situated at the surface of the transcripts. To find out how do the students talking (interacting)

with others, we adopt content analysis to investigate the possible process patterns within the collaboration interaction. That is, annotate the speech intention of discourse records and then make a quantitative analysis to discover the distribution of speech intentions and changing trend.

Although this research technique is often used, standards are not yet established [16]. The applied coding schemes reflect a wide variety of approaches and differ in their level of detail and the type of analysis categories used. Further differences are related to diversity in their theoretical base, the amount of information about validity and reliability, and the choice for the unit of analysis. So far, many researchers have proposed diverse coding schemes for content analysis. Henri proposes a coding scheme consists of five dimensions: participative, social, interactive, cognitive and metacognitive [5]. Newman et al. argue that there is a clear link between critical thinking, social interaction and deep learning, and accordingly developed a coding scheme composed of 10 categories [10]. Based on the combination of Vygotsky's theory and theories sof cognitive and constructive learning, Zhu divides the social interaction into vertical interaction and horizontal interaction [18]. The coding scheme developed by Veerman and Veldhuis-Diermanse [14] identifies two categories of messages: task-related and non-task-related messages. Task-related messages are further subdivided into three categories: new ideas, explanation, and evaluation. [2] presents the coding scheme for measuring cognitive, social, and teaching presence.

Nevertheless, Rourke and Anderson [12] suggest that instead of developing new coding schemes, researchers should use schemes that have been developed and used in previous research. Applying existing instruments fosters replicability and the validity of the instrument [13]. Therefore, we adopt the coding scheme developed by Chen-Chung Liu [8] to explore how learners collaboratively work on the task and formulate arguments together during collaborative interaction.

3.2 Theme-based Topic Recognition

Usually teachers or researchers are interested to know to what extent students' discussion overlap with expert's or textbook's conception on a certain themes in a discussion. So, it is useful to recognize the emerged topics in the discussion transcripts. Topic detection and tracking (TDT) research [1] [3] mainly focus on detecting and tracking events in streaming news data. TDT systems monitor continuously updated news stories and try to detect the first occurrence of a new story; i.e., an event significantly different from those news events seen before. Based on the approaches, text mining technology is adopted to discover the emerged topics in the discourse records of students. The key idea is that teachers initially present the themes that are expected to be talked by the students, and then the postings in a discussion thread are combined into a document. Afterwards, parse the documents and compute the semantic similarity with the theme vector proposed by the teachers. If the document is similar to an existing theme, the postings in the documents will be labeled with such theme; otherwise, it will be labeled as a new theme with related keywords description.

Assuming that the postings in a discussion thread represent the same topic, we combine them into a summary document and then process it. Each document is

represented as a weighted term vector $d = (d_1, d_2,\ldots)$ with the standard TFIDF function.

$$d_i = TF(w_i, d) \log(\frac{|E|}{DF(w_i)}) \quad (1)$$

Where the term frequency TF(w_i, d) is the number of times word w_i occurs in document d, $|E|$ denotes the total number of documents in the training set and the DF(w_i) is the number of postings containing the word w_i at least one time.

We consider the text in title field and body field of postings separately but discriminatively. Usually, title is the outline of body contents, so words in title filed are more descriptive and discriminative in contrast to the words in body field. Thus, words in title field are assigned larger weights to reinforce their stronger impact. For TF(w_i, d), one time appearance in title field equals to t times appearances in body field. The cosine method is adopted to compute the similarity between the document vector and the theme vector defined by the teachers, and thus the documents belong to the certain concept with the maximum similarity value.

3.3 Augmented Social Network Analysis

Social Network Analysis (SNA) is an established method to derive person-person relations in the form of sociograms from "traces" of communication in a networked community [15]. It is widely used to study the way people participated and interacted with each other in discussion boards [6], which provides information about the activities of such a community and the way they learn collaboratively. The discussion transcripts can be treated as relational data and stored away in a case-by-case matrix to analyze interaction patterns. A few of indicators are computed in SNA, such as betweenness, centrality, clique, cohesion, to indicate the activities of such a community and the way they learn collaboratively. But this method is simply based on the information flow between learners but ignore the content of postings, so the constructed social network is very large and complex. To better reflect the peer collaboration in CSCL, we focus on the theme-centered social network of the peers who are engaged in the same theme.

After determining the theme of each threaded-notes with above-mentioned method, the following formula are used to compute several criteria for evaluating a student's performance in the collaborative interaction, including participation, authority, novelty, coverage, and activity.

Participation:

$$P_i = \beta \sqrt{\frac{I_i}{\sum_{f=1..l} I_f}} + (1-\beta) \sqrt{\frac{R_i}{\sum_{f=1..l} R_f}}, \beta \in [0,1] \quad (2)$$

Where Ii denotes the number of postings initiated by the ith person, Ri denotes the number of postings replying to others posted by the ith person, l denotes the total number of students involved in the theme discussion.

Authority:

$$A_i = \frac{\sum_{t=1..I_i} \left(\frac{I_t^c}{max(I^c)} + \frac{I_t^r}{max(I^r)} \right)}{2I_i} \quad (3)$$

Where I_t^c denotes the times of clicking by others for the tth posting initiated by the ith person, I_t^r denotes the times of replying by others for the tth posting initiated by the ith person. Ii denotes the total number of postings delivered by the ith person. max(Ic) and max(Ir) denotes the maximum clicking-times and replying-times for a posting.

Novelty:

$$N_i = \frac{M_f}{M} \quad (4)$$

Where M_f represents the number of theme-related keywords mentioned for the first time by *i*th student, *M* represents the total number of keywords mentioned by the *i*th student.

Coverage:

$$C_i = \frac{I_i^k}{I_i} \quad (5)$$

Where I_i^k denotes the number of postings pasted by the *i*th student that belong to the *k*th theme, I_i denotes the total number of postings pasted by the *i*th student.

Activity:

$$AC_i = \frac{N_i}{\Delta t_p} \times \frac{1}{(t - t_d) + \tau} \quad (6)$$

Where N_i represents the number of postings posted by the *i*th person during the period Δt_p, t is the current date and t_d is the date the *i*th person posted the latest posting in the forum. τ is the adjust parameter to avoid the denominator is zero, and it is initially assigned 1.

4 Implementing a tool to support collaborative interaction analysis

We have developed a tool VINCA (Visual Intelligent Content Analyzer) with C# language to support interaction analysis. It is implemented by using C/S architecture and can be installed stand-alone or support the online downloading of the forum text from CSCL platform to conduct analysis. The tool provides a plug-in interface allowing for flexible addition of more modules. Figure 2 shows the framework to design the content analysis tool. It mainly comprises three modules: data preparation, text analysis, as well as visualization & Export. The preparation module allows the

users to import data in the format of HTML files, XML files, database, or text from different CSCL platforms, and then transforms the data into a standard relational database format automatically. The text analysis module is to analyze the raw transcripts or the coded transcripts with the support of keyword extraction, concordance viewing, and text similarity computation. The visualization & Export module provides visualization (e.g. graphs, curves, tables) of analysis results or export the multiple analysis results in the format of .csv files for further quantitative and code co-location explorations. More information can refer to [7].

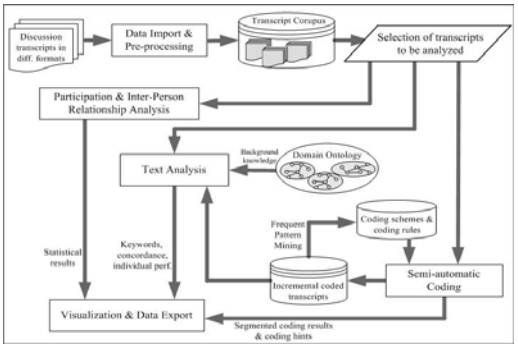

Fig. 2. Framework for designing the content analysis tool.

Figure 3 - figure 6 are the snapshots of the tool interface. Figure 3 shows the semi-automatic coding interface. As the figure shows, the coding hint is marked with red color and its corresponding recommended candidate codes are listed in the right part of the interface with the support percentage and confidence percentage. During the coding process, VINCA will scan each segment to locate the coding hints, highlight the hints, and then recommend the candidate codes with two computerized indicators: support and confidence percentage. The support indicator represents the hints appearance frequency in the transcripts corpus, while the confidence indicator means the reliability of the recommended codes. Users can accept the recommended codes or refuse it by selecting other code. The coding process and coding errors will be recorded. Thereafter, VINCA will make use of the hints, mistakes and missing lists of discourse segments and the coding effectiveness statistics to improve on the coding rules. Figure 4 illustrates the visualization of the coding results, including the coding distribution and coding changing trends. Figure 5 shows the interface to set the analysis parameter and show the extracted keywords, frequency, speakers, etc. The users can also click any keyword to view its concordance in the lower part of the interface. Furthermore, after importing the domain ontology constructed by the teachers or the researchers, VINCA support the evaluation of individual's performance by computing the topic relevance, novelty, and extension. One outstanding feature of VINCA is its coding rule learning mechanism by discovering the frequent pattern from the database of increased coding expertise. As figure 6 illustrates, the users can set the configuration for the pattern discovery and then check the resulting pattern in the lower part of the interface. Users can then select all or some patterns to add into the coding rules database.

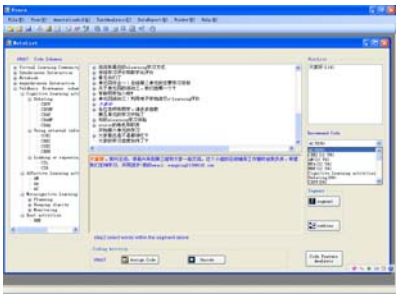

Fig. 3. Snapshot1

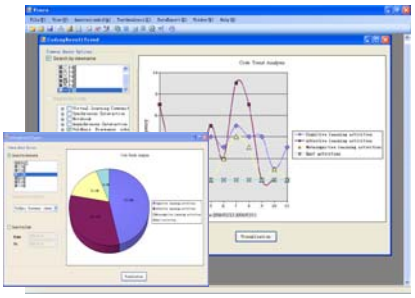

Fig. 4. Snapshot2

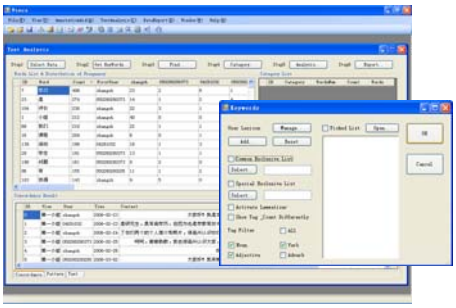

Fig. 5. Snapshot3.

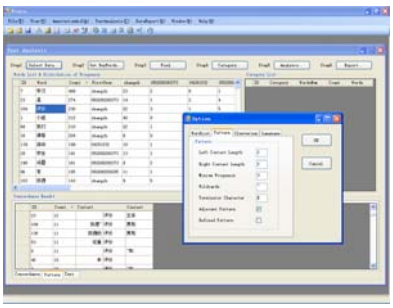

Fig. 6. Snapshot4.

5 A Case Study

We conduct an experiment to investigate the peer interaction in CSCL. A class of 18 graduate students majoring in Education Technology" enrolled in the course "Key Technologies in E-Learning and Application". During the semester from Sep. 2007 to Jan. 2008, the students took the course with a hybrid learning of two-lesson face-to-face learning each week as well as collaborative learning in Knowledge Forum (KF: http://kf.cite.hku.hk) anytime. Except the learning in the classroom, the graduate students were required to fulfill the assigned activities through online discussion. We chose a set of discourse data recorded in the KF platform as data source, and then use the coding component, text analysis, and data export component of VINCA to help analyzing the sampled data, for the purpose of unveiling the students' peer interaction in terms of process pattern, topic space and social network.

- **Process Pattern**

With the assistance of VINCA, two coders took the meaning unit as the basic analysis unit and performed the coding with the coding schema. The coding scheme is tabulated in table 1. After finishing the coding of discussion transcripts, users use the coding visualization module to view the coding distribution and coding changing trends. Figure 7 illustrates the coding results for each student. As the histogram shows, the student 8 performs better than other students with maximum notes. Furthermore, figure 8 displays the change trend of each type of notes during a fragment of the discussion period. Time sequence analysis of the discourses indicated that positions

often outnumbered issues, and issues were proposed and positioned increasingly during the initial stages of the activity. Following the initial stage, issues and positions decreased dramatically. Additionally, argument increased a lot in the middle stage, but decreased sharply after the middle age. Response keeps a relatively steady change trend during the whole process.

Table 1. Coding scheme.

Type	Meaning
Issue	What needs to be done and problems to solved, and related to the concepts and skills being learned by students.
Position	Methodologies for resolving an issue, and are answers from peers in response to issues that have been raised.
Argument	Opinions that support or object to a position
Group development	Questions raised to coordinate members to work together
Response	A suggested answer to a group development question
Acceptance of response	The acceptance or agreement of a response
Objection to response	Student objection or disagreement to responses
Conflict	Contradiction occurs among students
Support request	A request for resources and help from other group members

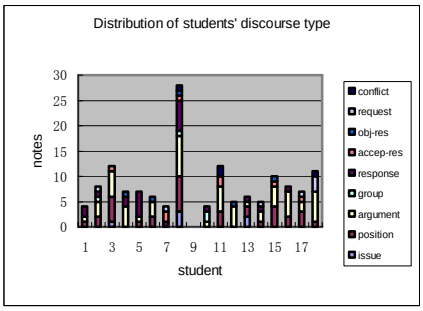

Fig. 7. Visualization of coding statistics

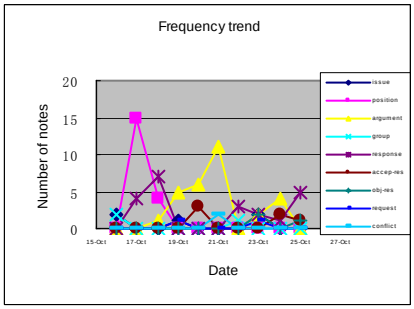

Fig. 8. Changing trend of each type of notes.

Additionally, students were divided into five groups and were assigned an activity of "Design of vertical search engine". Each group was asked to collaboratively determine the topic of the search and task allocation through online discussion in knowledge Forum. Afterwards, their discourse records were analyzed via using VINCA to further investigate the interaction pattern among members. Herein we select two groups (group A and group B) for illustrative purpose. Figure 9 shows the interaction pattern graphs for the two groups, respectively. As the figure indicates, two groups has quite different interaction pattern. Regarding group A, a2 plays a central role within the group by organizing the group collaboration and receiving many responses from other members. This kind of interaction can be defined as centralized knowledge exchange. By contrast, there is no central member within

group B. All of the members in group B interact with each other by expressing opinions or giving answers to other's question. So, this kind of interaction is more likely to be called distributive knowledge exchange.

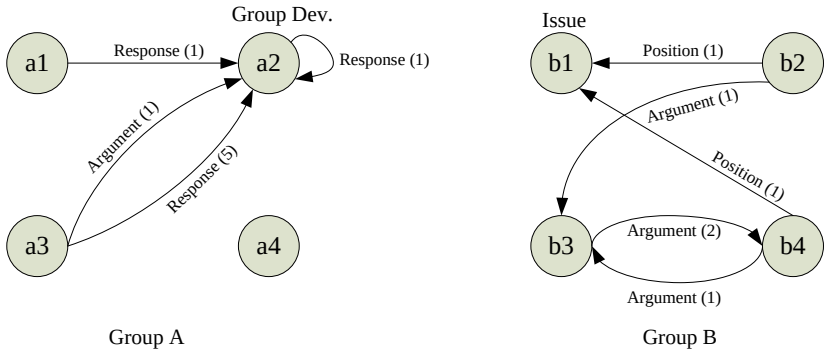

Fig. 9. Peer interaction within two groups.

- **Topic Space**

A learning activity assigned by the teacher is "discussing on the web-based course analyzing", and all the students were required to exchange their ideas via the KF platform. Regarding this activity, the teacher assigned two themes, including "web-based course evaluation" and "web-based course design". After computing the similarity of each threaded-notes to the themes by using the text analysis module of VINCA, the resulting topic space consists of another new discovered themes, including "learner characteristics", "teaching effect", "perfect course", "person of ability". Figure 10 intuitively shows the constructed topic space for the students' discussion. As the figure shows, there are in total 6 topics, and t4, t5 involve more students' discussion compared to other topics.

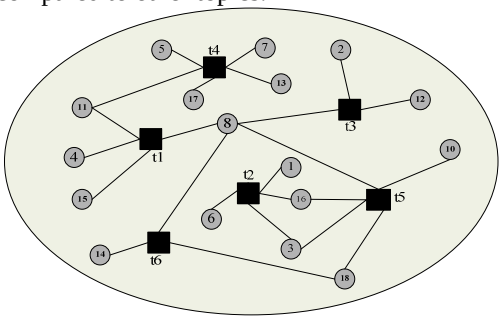

Fig. 10. Topic space.

- **Social Network**

Different from the traditional social network analysis, we herein focus on the analysis of the social relationship between students who are involved in the discussion on the same theme. Figure 11 shows the social network on the theme "perfect course". From the figure, we can see that 5 students participated in the discussion with active interaction among them. The wider the edge between the students is, the more interactions between them occur. S16 plays a central role in the discussion by drawing

attentions from other students and especially S18 responses to S16 a lot. To further illustrate the learner characteristics in the collaborative learning, we compute the several indicators for the selected two students: S8 and S16. As figure 12 shows, S8 has higher level of participation and activity compared to S16, but his other characteristics such as authority, novelty, and coverage are relatively lower than S16. It implies that though S8 are very active engaged in the collaborative learning, his speeches do not draw a lot of attention from others, whilst S16 attains others' more responses with relatively small speeches.

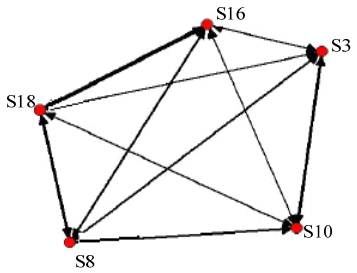

Fig. 11. Theme-centered social network

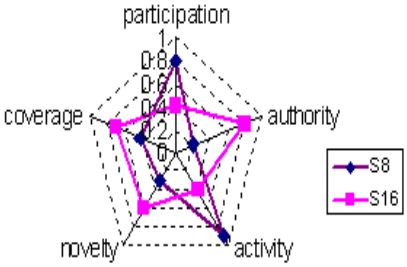

Fig. 12. Learner characteristics

6 Conclusions

By incorporating content analysis, social network analysis, and data mining technologies, this paper puts forward a three-dimensional model to help researchers understand what happened behind online peer interactions in CSCL. This paper also designs and implements an intelligent content analysis tool VINCA to support analyzing peer discussion transcripts. An experimental study is conducted to explore the discussion topics, process pattern and social network with respect to the students' online interaction, while illustrating the viability and usefulness of VINCA during the analyzing process.

References

1. J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In 26th annual ACM SIGIR Conference, Toronto, Canada, July 2003.
2. T. Anderson, L. Rourke, D. R. Garrison, & W. Archer, (2001). Assessing teaching presence in a computer conference context. Journal of Asynchronous Learning Networks, http://www.sloan-c.org/publications/jaln/v5n2/pdf/v5n2_anderson.pdf (Retrieved: August 15, 2004).
3. T. Brants, F. Chen, and A. Farahat. A system for new event detection. In 26th annual ACM SIGIR Conference, Toronto, Canada, July 2003.
4. P. Hakkinen, S. Jarvela, K. Makitalo, Sharing perspective in virtual interaction: Review of methods of analysis, In B. Wasson, S. Ludvigsen & U. Hoppe (Eds.), Designing for Change in Networked Learning Environments, proceedings of the International Conference on Computer-support for Collaborative Learning 2003, pp.395-404, Dordrecht: Kluwer.

5. F. Henri, Computer conferencing and content analysis. In A. R. Kaye (Ed.), Collaborative learning through computer conferencing. The Najadan Papers, pp. 117–136, (1992).
6. M. D. Laat, V. Lally, Investing group structures in CSCL: some new approaches, Information Systems Frontiers 7:1, 13–25, 2005.
7. Y. Li, J. Wang, J. Liao, D. Zhao, R. Huang, Assessing Collaborative Process in CSCL with an Intelligent Content Analysis Toolkit, IEEE International Conference on Advanced Learning Technologies (IEEE ICALT), Japan, 2007, pp. 257-261.
8. C. C. Liu, C. C. Tsai, An analysis of peer interaction patterns as discoursed by on-line small group problem-solving activity, Computers & Education, 2006
9. K.Meyer, (2004). Evaluating online discussions: four different frames of analysis. Journal of Asynchronous Learning Networks, 8(2), 101–114.
10. D. R. Newman, B.Webb, & C. Cochrane, (1995). A content analysis method to measure critical thinking in face-toface and computer supported group learning. Interpersonal Computing and Technology, 3, 56–77. http://www.qub.ac.uk/mgt/papers/methods/contpap.html (Retrieved August 15, 2004).
11. K. Nurmela, T. Palonen, E.Lehtine, K. Hakkarinen, Developing tools for analyzing CSCL process, proceedings of the International Conference on Computer-support for Collaborative Learning 2003, Bergen, Norway, 2003, pp.333-342.
12. L.Rourke, T.Anderson, (2003). Validity in quantitative content analysis. (Retrieved August 1, 2004), from http://communitiesofinquiry.com/sub/papers.html
13. E.Stacey, P.Gerbic, (2003). Investigating the impact of computer conferencing: content analysis as a manageable research tool. In G. Crisp, D. Thiele, I. Scholten, S. Barker, & J. Baron (Eds.), Interact, integrate, impact: Proceedings of the 20th annual conference of the australasian society for computers in learning in tertiary education, Adelaide, 7–10 December 2003. (Retrieved September 1, 2004), from http://www.ascilite.org.au/conferences/adelaide03/docs/pdf/495.pdf.
14. Veerman, E.Veldhuis-Diermanse, (2001). Collaborative learning through computer-mediated communication in academic education. In Euro CSCL 2001 (pp. 625–632). Maastricht: McLuhan institute, University of Maastricht.
15. S. Wassermann, K. Faust, Social Network Analysis: Methods and Application. Cambridge University Press, Cambridge, 1994.
16. B. D. Wever, T. Schellens, M. Valcke, H. V. Keer, Content analysis schemes to analyze transcripts of online asynchronous discussion groups: a review, Computers & Education, 46(1), 2006, pp.6-28.
17. B.,D.Wever, T.Schellens, M. Valcke, (2004). Samenwerkend leren via informatie- en communicatietechnologie [Collaborative learning through ICT]. In I. D'haese & M. Valcke (Eds.), Digitaal leren. ICT Toepassingen in hethoger onderwijs [Digital Learning. ICT applications in higher education]. Tielt: Lannoo Campus.
18. E.Zhu,. Meaning negotiation, knowledge construction, and mentoring in a distance learning course. In: Proceedings of selected research and development presentations at the 1996 national convention of the association for educational communications and technology. Indeanapolis: Available from ERIC documents: ED 397 849., 1996.