2009 Ninth International Conference on Intelligent Systems Design and Applications

Analyzing Online Asynchronous Discussion Using Content and Social Network Analysis

Erlin
Department of
Information Technology
STMIK-AMIK Riau, 28294
Pekanbaru, Indonesia
erlin_by@yahoo.com

Norazah Yusof
Department of
Software Engineering
Universiti Teknologi Malaysia
81310 Skudai, Malaysia
norazah@utm.my

Azizah Abdul Rahman
Department of
Information Systems
Universiti Teknologi Malaysia
81310 Skudai, Malaysia
azizahar@utm.my

Abstract—**Asynchronous discussion forum can provide a platform for online learners to communicate with one another easily, without the constraint of place and time. This study explores the analysis process of online asynchronous discussion. We focus upon content analysis and social network analysis, which is the technique often used to measure online discussion in formal educational settings. In addition, Soller's model for content analysis was developed and employed to qualitatively analyze the online discussion. We also discuss the use of network indicators of social network analysis to assess level participation and communication structure throughout online discussion. Adjacency matrix, graph theory and network analysis techniques were applied to quantitatively define the networks interaction among students. The findings showed that these methods provide more meaningful students' interaction analysis in term of information of communication transcripts and communication structures in online asynchronous discussion.**

Keywords-online discussion; content analysis; social network analysis; centrality; density; reliability; coding categories

I. INTRODUCTION

In this information age, the application of information and communications technology (ICT) for learning has become progressively more popular, mainly due to the many assumed benefits of the use of computer-mediated communication (CMC). Similarly, the increasing popularity of the learning technology and internet and its ability to provide seemingly transparent communication between different computing platforms has simplified the processes of providing learning opportunities to remotely located learners.

CMC refers to communication between individuals and among groups via networked computers [1]. Such form of communication can be either asynchronous or synchronous and serve a wide variety of useful functions. De Wever in [2] found that asynchronous discussion forum take a central place in CMC environment.

An increasing number of researchers have attempted to produce techniques that measure and analyze quality of asynchronous discussion. Spatariu in [3], having reviewed current literature, suggest that the majority of studies can be loosely categorized into one of four categories, according to the construct being measured; levels of disagreement, argument structure analysis, interaction based and content analysis.

Although, many researchers have argued that content analysis and social network are the popular methods to analyze asynchronous discussion [4][5][6][7] there has not been much exploration of integration of these methods. Insight in collaborative learning within a certain analysis transcripts alone is not sufficient. Also the analysis of the network must be taken in account. This allow researchers to examine the phenomena from different perspective because in the past, the majority of study in CMC using this methods in separate way.

Therefore, this study employs content analysis and social network techniques to analyze asynchronous discussion for students participating in a course. The quality of the content of communication is evaluated through content analysis; and the network structures are analyzed using social network analysis of the response relations among students during discussions. Thus, using content analysis and social network analysis offer a solution to analyze the transcripts and network structures for measuring quality and quantity of asynchronous discussion forum by providing the best of both methodologies.

II. DESCRIPTION OF THE METHODS

A. Content Analysis (CA)

We have explored fundamental issues in content analysis; unit of analysis, reliability and coding scheme. Choice the unit of analysis is the starting point for coding the transcripts. Rourke et al. in [8] distinguishes five types of units, from large to small, a message, a paragraph, unit of meaning, sentence and illocution. The most frequently reported units are a message, a unit of meaning and the sentence [9].

Reliability in the context of content analysis refers to the amount of agreement or correspondence among two or more coders [4]. The reliability computed as the proportion agreement because there is only one category involved with two values (agree=1, disagree=0); two or more categories

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.40

requires computation of one or multiple coefficient to correct for chance agreement.

Many researches have constructed coding schemes or model for what they want to explore in the content and process of online discussion. From a cognitivist viewpoint, Henri in [10] has developed a model to analyze the transcripts of discussions. Her framework has five dimensions: participative, social, interactive, cognitive and meta-cognitive. She made operational definitions of each of the dimension. Gunawardena in [11] has also developed a model, based on grounded theory. They used the phases of a discussion to determine the amount of knowledge constructed within the discussions analyzed. Moreover, a collaborative learning skill category was developed by Soller [12], required students to use a given set of sentence openers and each sentence opener is associated with a particular conversational intention, given by three main category variables and nine sub category variables. Her model using sentence opener approach in which the coding category is manually chosen by students.

We adapted Soller's model as our research tool because her model provide a wide range of analytical dimensions which can best support our research purpose. Additionally, we modified Soller's model so that our data can be analyzed most appropriately.

B. Social Network Analysis (SNA)

We have identified a set of SNA indicators for the study of participatory aspects of learning; degree centrality, density and network degree centralization [13]. Degree centrality is the degree of each actor. It is a method of evaluating centrality on the basis of a student's direct linkage to other students. In a directed network that considers the direction of the link, two degree centrality is presented by in-degree centrality and out-degree centrality. In-degree centrality means the number of the links terminating at the node and out-degree centrality, on the other hand, means the number of links originating at the node.

Density provides a measure of the overall connections between the students. This gives an indication of the level of engagement in the network. Density calculations indicate how active the students are involved in the discourse [6]. The density of a network is defined as the number of communicative links observed in a network divided by the maximum number of possible links [14]. This varies between 0 and 1. When the density is 0, the network is without any connection; and when the density is 1, all the students of a network are connected to one another.

Finally, network degree centralization is a group-level measure based on actor's degree centrality. Directed networks define the corresponding indexes of in-degree centralization and out-degree centralization. All of these indexes and ranges apply to dichotomous relationships that can have only one out of two possible values: 0 when there is no link and 1 when there is a link between two actors.

III. METHODOLOGY

In order to pilot test the efficacy of CA and SNA for analyzing online asynchronous discussion, the selected transcripts of subject SCJ2013-01 2008/2009: Data Structure and Algorithm held on Moodle as a learning management system (LMS) in e-learning was examined for one thread. This thread was chosen because it has highest replies than other threads. There were 12 students completed the thread discussion.

A. Data Collection

Data for this study were transcripts of students' discussion using the threaded discussion tool on Moodle. The total number of messages in the discussion forum (that were replies to somebody's message) was 38 (excluding 1 message from initiator). The transcripts were important to analyze the dynamics of forum and what kind of feedback from one another.

B. Data Analysis

The qualitative data were analyzed using content analysis and choose sentence as a single unit of meaning and would be validate by two coders to calculate reliability of segmentation and coding categories. Moreover, adjacency matrix, graph theory and network analysis technique were applied to quantitatively define the network interaction among student. The data were saved in matrix for analysis purposes.

C. Procedure

The procedure for integrating CA and SNA would be started on procedure for CMC content analysis. It should comprise, at least, five steps [9]. First step is determination of unit of analysis. This study selected sentence as a single unit of meaning due to sentence is closer to interpret as a unit of analysis.

Second step is development of segmentation procedure to break message into sentences. The transcripts segmentation component allows the users to segment the text into sentences. Each message is first segmented in sentences by using full stop, question mark or exclamation mark that the author of the message has written (except only one word). A total of 38 messages in this threaded would be split into 95 sentences. The coders agree in 87 agreements which is 81 agree as sentences and 6 agree as not sentences. Disagree in 8 sentences.

The reliability test was conducted by two coders on segmentation procedure using multiple reliability coefficients such as percent agreement, Scott's Pi (π), Cohen's Kappa (κ) and Krippendorff's alpha (α). Reporting multiple reliability indices is of importance considering the fact that no unambiguous standards are available to judge reliability values. The coders do a sample exercise on other messages to familiarize themselves with the model. Two coders should do the analysis independently and have the results cross examined by one another.

The reliability value of segmentation procedure is shown in table 1 as follows: PA=91.6%, π =0.726; κ= 0.727 and α =0.727. Krippendorff added that variable with Alpha as low as .667 could be acceptable for drawing tentative conclusions [15]. The values of .667 also appropriate for Scott Pi and Cohen Kappa.

TABLE I. THE RELIABILITY OF SEGMENTATION PROCEDURE

N Agreement	87
N Disagreement	8
N Cases	95
N Decisions	190

Percent Agreement	Scott Pi	Cohen's Kappa	Krippendorff Alpha
91.6%	0.726	0.727	0.727

Next step is development of coding categories. We adapted and developed the coding categories based on Soller's model as shown in table 2.

TABLE II. CODING CATEGORIES

Code	Category
Creative Conflict : Mediate	
11	Recommended an instructor intervene to answer a question
Creative Conflict : Discuss	
12	Discuss and give a reason (positively or negatively) about comments or suggestions made by team members
Active Learning : Motivate	
23	Providing positive feedback and reinforcement
Active Learning : Inform	
24	Direct or advance the conversation by providing information or advice
Active Learning : Request	
25	Ask for help/advice in solving the problem, or in understanding a team-mates comment
Conversation : Acknowledge	
36	Inform peers that you read and/or appreciate their comments. Answer yes/no questions
Conversation : Maintenance	
37	Support group cohesion and peer involvement
Conversation : Task	
38	Shift the current focus of the group to a new subtask or tool
Non Codable	
40	All types of statements that not belong to any category specified (e.g., statements that signal receipt of a message or attachment)

After the initial round of coding, the next step is determination of reliability of coding categories. The value of reliability of coding categories as follows: Percent Agreement=93.8%, π =0.919; κ= 0.919 and α =0.920. These values concluded that coding categories could be acceptable for drawing conclusion.

For processing into SNA, the number of sentence breakdown from message which is segmented during content analysis process would be taken as an input to SNA. We applied sentence instead of message because it was more fair to analyze student's interaction. Message may consist of several sentences but only counted as one unit interaction. Further, the data can be treated as relational data and stored away in matrix to analyze interaction patterns.

We conducted centrality measures to find the central students within the network. The network activity of individual members can be indicated. This can be done by calculating the in-degree centrality and out-degree centrality measures. Secondly, we conducted a density analysis to describe the overall linkage between students in online discussion. Density can show how dense is the participation within it. Finally, network degree centralization would be calculated to perform a group level measure based on actor's centrality. It gives illustrate in the dependency of the network on the activity of group of actors.

The network analysis software-NetMiner [16] was used to conduct the analysis and represent these online interactions in visual object.

IV. RESULT AND DISCUSSION

Altogether we analyzed 38 messages containing 95 sentences, which were posted by 12 students. The discussion was started by student A. Only 81 sentences as an agreement result from two coders during segmentation would be process for content analysis and SNA.

A. Students' Interaction in Content Analysis (CA)

TABLE III. CATEGORY STATISTIC

SID	Creative Conflict		Active Learning			Conversation			NC	D's	Σ
	11	*12*	*23*	*24*	*25*	*36*	*37*	*38*	*40*	*D's*	
A	0	4	1	4	3	4	0	0	0	1	17
B	0	1	0	0	1	0	0	0	0	0	2
C	0	0	0	0	1	0	0	1	0	0	2
D	0	0	1	2	0	0	0	0	0	0	3
E	0	0	0	2	0	2	0	0	1	1	6
F	0	0	0	0	0	0	0	1	0	0	1
G	0	2	0	1	0	0	0	0	0	0	3
H	0	0	1	1	5	3	0	0	0	3	13
I	0	2	0	3	1	0	0	1	0	0	7
J	0	1	0	12	0	0	0	2	0	0	15
K	0	0	2	2	0	0	1	0	0	0	5
L	0	1	0	4	2	0	0	0	0	0	7
∑	0	11	5	31	13	9	1	5	1	5	81

*SID = Student's ID, NC=Non codable, D's = Disagreements coding

The results given in table 3 showed that relationship between student and each of categories. The sentences for which no category was reached as to main category were coded in a non codable (NC). The sentences for which no agreement was reached any coders were coded in disagreement (D's). Interestingly, this table shows no one student play a role in mediate sub category. Student who is more used several category are student A, student J and student K; 17, 15, 13 respectively. Each student is minimum posting one message in this discussion.

Table 4 shows the final statistics for the thread at the end of discussion for this threaded. The most frequently involved interaction type in main category was 'active learning skills' (60.49%) and the most frequently used sub category was 'inform' (38.27%). Only 13.58% of the ideas revealed creative conflict skills (i.e., those of mediate and discuss), 18.52% of ideas revealed conversation skill (i.e., acknowledge, maintenance and task), 1.23% of ideas can not approximated for one category or non codable and 6.17% of the ideas for which no agreement was threaded by two coders as disagreement code.

TABLE IV. THE PERCENTAGE OF SENTENCES BY SUBCATEGORIES WHICH MAKE UP EACH CATEGORY TYPE

Main Category	Percentage
Creative Conflict (11+12)	13.58%
Active Learning (23+24+25)	60.49%
Conversation (36+37+38)	18.52%
Non Codable (40)	1.23%
D's (Disagreement Coding)	6.17%

Sub Category		Percentage	
Mediate	Creative Conflict	0	13.58%
Discuss		13.58%	
Motivate	Active Learning	6.17%	60.49%
Inform		38.27%	
Request		16.05%	
Acknowledge	Conversation	11.11%	18.52%
Maintenance		1.23%	
Task		6.17%	

Example of result that can be obtained from the online discussion is shown in fig. 1 that shows data from twelve students who naturally played different roles during discussion.

Qualitative analysis of the transcript shows that the student J played the role of an advisor or informer, making specific recommendations and providing information to the others. Student H was significantly perform more request compare to other students. He/She played a questioner, asking several clarification questions. Student K played the role of a motivator, giving and providing positive feedback and reinforcement to others in their group. This supports the idea that a student's role may be partly, or fully, determined by the types of category acts he/she was using.

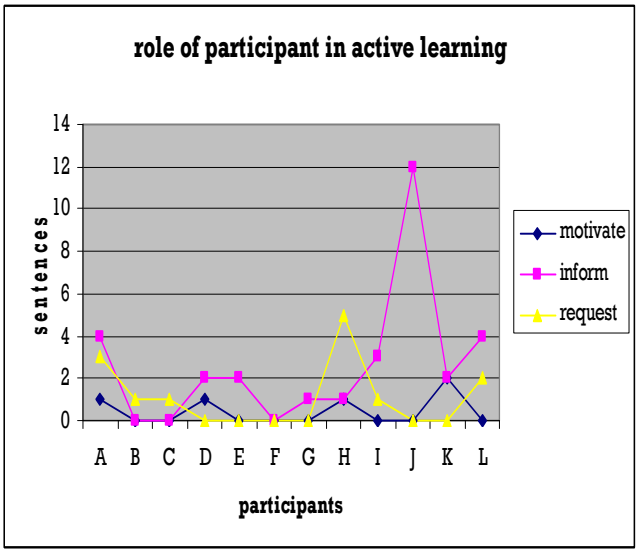

Figure 1. ROLE of STUDENTS in ACTIVE LEARNING

B. *Students' Interaction in Social Network Analysis (SNA)*

In table 5, the resulting adjacency matrix of interactions in the online asynchronous space is shown that showed which students responded and posts to each other and how often they did so. The initial posts were not considered interaction and were not counted. This is cumulative data: for example, it shows that the student A responded to student H four times and to student J five times.

TABLE V. ADJACENCY MATRIX

↗	A	B	C	D	E	F	G	H	I	J	K	L
A								4		5	4	4
B	2											
C	2											
D	3											
E				1						5		
F	1											
G	3											
H	5								3	5		
I	3					1		3				
J	10							2	3			
K	5											
L	4	2				1						

The data was imported into the network analysis software-NetMiner to analyze the interaction among student. Centrality measures are being conducted to find central actors in a network. This can be done by calculating the in-degree and out-degree measure (table 6). Student A is high in-degree indicates that he/she receives more information or comments from others and this student has more prestige in the network. Unfortunately, there were three student have in-degree value is 0 meaning that they did not get any replies from others. Student A, also high in out-degree indicates that he/she is more active in providing information to others or providing comments and opinions of others. Some of the students, on the other hand, low in out-degree meaning that they hardly participated within the discussion at all.

TABLE VI. IN-DEGREE AND OUT-DEGREE

		1	2
		In-Degree	Out-Degree
1	A	38.000000	17.000000
2	B	2.000000	2.000000
3	C	0.000000	2.000000
4	D	1.000000	3.000000
5	E	0.000000	6.000000
6	F	2.000000	1.000000
7	G	0.000000	3.000000
8	H	9.000000	13.000000
9	I	6.000000	7.000000
10	J	15.000000	15.000000
11	K	4.000000	5.000000
12	L	4.000000	7.000000

Student A was the only actor who ranked higher in both in-degree and out-degree considered so far. This student was prolific and consulted very often. However, the student that had lower rates in both dimension can be classified as lurkers or isolates.

TABLE VII. REPORT ON IN-DEGREE AND OUT-DEGREE

Measures	Value	
	In-degree	Out-degree
Sum	81	81
Mean	6.75	6.75
Std. Dev.	10.329	5.182
Min	0	1
Max	38	17
# of isolate	0	
Network Density	0.614	

In table 7 report on in-degree and out-degree are shown. To get an indication of the overall linkage of students in the network we conducted density calculations that indicate how active the students are involved in the discussion and show how dense is the participation within it. In this case of sending and receiving the sentences that were exchange through online discussion had a density of 61.4% within 81 sentences. Student in-degree varied between 0 and 38 and out-degree varied between 1 and 17.

Fig. 2 presents graph of out-degree centrality. It is clear from the graph that two high extremes are measured for 2 students (A and J), they have an out-degree centrality of 1.545 and 1.364 respectively. They are the most powerful actors of the network and they are positioned toward the center of the out-degree centrality circle. They actively participate and provide information and comments on the opinions of others. They also have friendly relations with many students and have important roles in delivering information to their community. The less powerful students or lurkers were student B, C, D, F and G. Based on this graph, instructors can trigger them to more active in providing or delivering information to others.

Figure 2. GRAPH of OUT-DEGREE CENTRALITY

Fig. 3 illustrates the mapping of interaction between twelve students in a directed graph. From the graph it can be said that student A get many replies and highest in term of in-degree centrality. Student C, student G and student K interact only with student A. Moreover, Student C, student F and student G interact in one way, meaning that they had been isolate from others. From this graph instructors can detect who is involved with the discussion, who is active student and who is lurker.

To obtain an indication of the dependency of the network on the activity of students, we conducted network degree centralization index. This gives illustration of the

level measure based on student's degree centrality. The stress values, that indicates the dependency of this discussion had a network out-degree centralization index of 309.917% and in-degree centralization index of 101.653%.

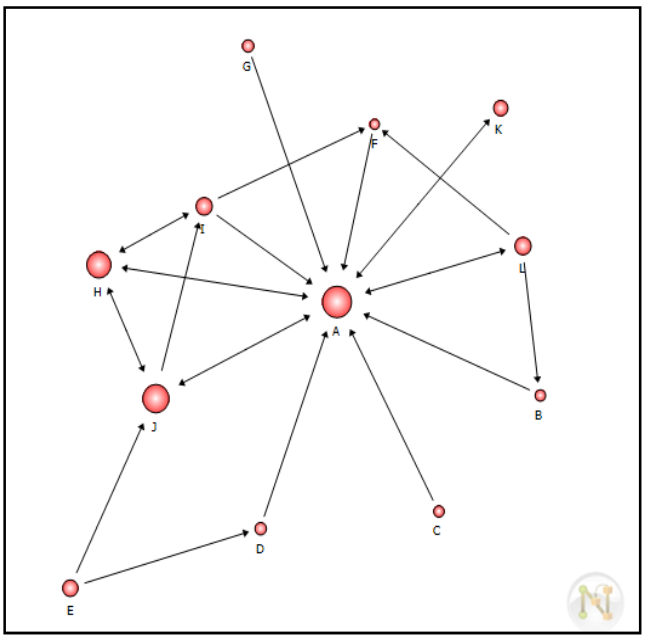

Figure 3. MAPPING of INTERACTION

V. CONCLUSION AND FUTURE WORK

We applied content analysis to categorize data into meaningful category and social network analysis to visualize communication structures.

In the content analysis, we examined three main category variables and nine sub category variables within message posting. Content analysis revealed that the most frequently involved interaction type in main category was 'active learning' and the most frequently used sub category was 'inform'. The reliability was conducted into two parts; segmentation procedure and coding categories by two coders. The values of reliability concluded that segmentation procedure and coding categories could be acceptable for drawing conclusion.

Moreover, in social network analysis, the visual objects that represent these online interactions are demonstrated and explained. SNA provided useful information about virtual interactions; information regarding the communication structures, level of participation, identifying who is central actor, who is involved in online discussion, bridge and isolate (able to determine who are not engaged in the discussion).

We are also aware of many challenges facing research of this nature. For instance, among others, manually doing content analysis is labour intensive and takes a long time. There are also issues of reliability which are difficult to overcome in situations where the coding scheme is emergent. In future, sentences segmentation and coding categories using a neural network method will be used to break a message into sentences to faster segmentation as well as for coding categories purposes.

REFERENCES

[1] Naidu, S and Jarvela, S, 2006, "Analyzing CMC content for what?", *Computer & Education* 46, pp. 96-103.

[2] De Wever, B., Schellens, T., Valcke, M. and Van Keer, H, 2005, "Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review", *Computers & Education* 46 (1), pp. 6–28.

[3] Spatariu, A., Hartley, K. and Bendixen, L.D. 2004, "Defining and measuring quality in on-line discussion", *Journal of Interactive Online Learning*. Vol 2 (4).

[4] Neuendorf, K. A., 2002, *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications.

[5] Anderson, T., Rourke, L., Garrison, D. R., and Archer, W, 2001, "Assessing teaching presence in a computer conference context", *Asynchronous Learning Networks*, from http://www.sloan-org/publications/jaln/v5n2/pdf/v5n2_anderson.pdf

[6] de Laat, M., 2002, "Network and content in an online community discourse", from http://www.uu.nl/uupublish/content/2002%20Networked%20Learning%201.pd

[7] Willging, P. A., 2005, "Using social network analysis techniques to examine online interactions", *US-China Education Review*, Vol 2, No.9 (Serial No.10).

[8] Rourke, L, Anderson, T., Garrison, DR and Archer, W., 2001, "Methodological issues in the content analysis of computer conference transcripts", *Artificial Intelligence in Education* 12**,** pp. 8–22.

[9] Strijbos, J. W., Martens, R., Prins, J., and Jochems, W., 2006, "Content analysis: What are they talking about?", *Computer & Education*, Elsevier, 46. 29-48.

[10] Henri, F., 1992, "*Computer conferencing and content analysis*. In A. R. Kaye (Ed.), *Collaborative learning through computer conferencing"*, London: Springer.

[11] Gunawardena, C. N., Lowe, C. A., and Anderson, T., 1997, "Analysis of global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing". *Journal of Educational Computing Research*, 17(4), 397–431.

[12] Soller A, 2004, "Computational modeling and analysis of knowledge sharing in collaborative distance learning", *User Modeling and User-Adapted Interaction* 14: 351–381.

[13] Wasserman, S., and Faust, K, 1997, *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

[14] Scott, J.**,** 2001, *Social Network Analysis: A Handbook*, 2nd ed., London: Sage.

[15] Krippendorff, 2004, *Quantitative content analysis: An introduction to its method*, Beverly Hills: Sage Publications.

[16] http://www.netminer.com/