Online MDL-Markov analysis of a discussion process in CSCL

Maomi Ueno and Toshio Okamoto
Graduate School of Information Systems, The University of Electro-Communications
E-mail: ueno@kjs.nagaokaut.ac.jp

Abstract

An online visualization system of a discussion process in computer-supported collaborative learning (CSCL) has been developed to assist learners by enabling them to monitor the actual states of their discussion and allowing them to improve their learning community. The unique features of this system are 1) the learners have to select the most suitable category that represents his/her message content from the choices presented in the developed BBS and 2) the proposed system estimates the structure of the Markov from the stored categories' data sequences and visualizes the structure online. We demonstrate the effectiveness of this system using actual data and provide some evaluations. The results show that the proposed system motivates the learners and improves their methods of discussion.

1. Introduction

The study of computer-supported collaborative learning (CSCL) is a challenge with regard to producing an environment conducive to mutual learning among learners using computers. Recent research in e-learning has highlighted the significance of building an online learning community, which plays a role in sustaining a fruitful online learning experience [1]. The significance of promoting communication among learners via computer-mediated communication (CMC) is rapidly increasing.

However, learners face some difficulties in mutually recognizing the status of a learning activity in the CSCL environment. This problem constitutes the most important research issue [2]. Kimura and Tsuzuki [3] pointed out that group communication in CMC tends to be disorganized and lack cohesion due to decreased interpersonal pressure, given the nature of CMC. To address this problem, we propose an online visualization system of discussion processes in an electronic forum to support learners in monitoring the condition of their discussion and thus improve their learning community.

Several recent studies in CSCL have focused on visualization of learner activities in CSCL in order to create awareness among learners. For example, Nakahara et al. [4] developed a software program that visualized the status of interaction and activeness of electronic forums on a mobile phone screen to promote participation awareness and encourage learners to participate in the discussion at any time. Martínez et al [5] have attempted to visualize social networks in the community, e.g., by confirming the status of communities in CSCL. Puntambekar and Luckin [6] have indicated that allowing learners to view the contents of the discussion and learn through reflecting over the process would be worthwhile. Mochizuki et al. [7] proposed a keyword analysis method using corresponding analysis, which is a text-mining technique, to assess conversation among learners on an electronic forum. Furthermore, they developed a visualization system based on [7] using the metaphor of the relationship between flowers and bees [8]. Ueno [9] also proposed a keyword-mapping method using extended corresponding analysis to mine some important keyword functions.

These previous studies focused on the visualization of the discussion content, but did not address the visualization of the discussion processes. The visualization of discussion processes in CSCL is expected to help learners, facilitators, and teachers to reflect the current discussion processes, thereby improving them. The system we propose focuses on the visualization of the discussion processes and uses a unique method that treats each message in the electronic forum as a code and assumes that the discussion process, which is defined by a series of message categories, follows a Markov source from the information theory [9] approach. The system then infers the structure of the Markov source using the minimum description length (MDL) from the data and visualizes the structure online. The advantages of the method are 1) that it provides optimal data compression and 2) that it has a strong consistency, which maximizes the prediction efficiency. We

Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT'06)
0-7695-2632-2/06 $20.00 © 2006 IEEE

have used some real data to show the effectiveness of this system.

2. Discussion process data

One of the authors has developed the LMS "Samurai" (see, for example, [9]). This system also has an electronic forum as well as other systems. One of the unique features of this electronic forum is that a learner has to select the most suitable category that represents his/her message content from the message categories shown when he/she contributes a message (Figure 1). The message categories are:

1. An opinion concerning the lecture content
2. A new opinion
3. Evidence that supports the previous message
4. Evidence that does not support the previous message
5. A consenting opinion on the previous message
6. A counterargument to the previous message
7. Caution to the previous message
8. A question about the previous message
9. A new question
10. A reply to the previous question
11. Indication of mistakes in the previous message
12. A digression
13. An encouragement to the previous message
14. An instruction from the teacher

The system stores the data sequence of these categories as a time series.. This data summarizes the large amount of data of the discussion process. The main issue we are dealing with is how to effectively utilize this data.

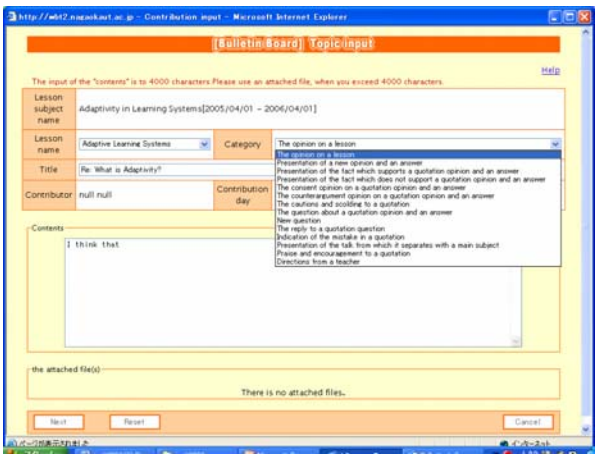

Figure 1. A category selection on the input message page

3. A mathematical model of discussion processes from information theory

3.1. Discussion process as a Markov process

The remaining problem is how to model this data. We used an information theory approach [10] to make a discussion process model. In this approach, each message in a discussion process is considered as a "code," and a series of the codes follows a Markov process.

The state c_k at time k is one of a finite number in the range $\{1,\cdots,M\}$. Under the assumption that the process runs only from time 0 to time N and that the initial and final states are known, the state sequence is then represented by a finite vector $C = (c_0,...,c_N)$. Let $P(c_k \mid c_0,c_1,...,c_{(k-1)})$ denote the probability (chance of occurrence) of the state c_k at time k conditioned on all states up to time $k-1$. Suppose a process was such that c_k depended only on the previous *n-th* states and was independent of all other previous states. This process would be known as an *n*th-order Markov process. That is,

$$P(c_k \mid c_0, c_1, \cdots, c_{k-1}) = P(c_k \mid c_{k-n}, \cdots, c_{k-1})$$

Although there are some solutions to estimate the probability structure of this model, it is well known that MDL is the optimal method. The next section introduces MDL.

3.2. Minimum description length (MDL)

The MDL is a statistical inference principle based on information theory [11], [12]. It says that among various possible stochastic models (or model classes) for a data sequence $c^k = c_1 \cdots c_k$, one should select the model yielding the shortest code for c^k, taking into account also the bits needed to describe the model (model class) that has been used for the encoding. The MDL has naturally led to a strong interplay with statistics of the theory of universal data compression in information theory. For binary sequences c^k, consider the model classes "i.i.d.," "first order Markov," "second order Markov," ..., "*n-th* order Markov." A code may be regarded optimal for a model class if the maximum over the model class of either its mean-redundancy or its max-redundancy is the smallest possible. Information theory has made it known that a code $f_n : \{0,1\}^k \rightarrow \{0,1\}^*$, optimal in either sense for the "order k Markov" model class (k=0 meaning

i.i.d.), has codeword length $\ell(f_n(x^k)) = -\log_2 P^k(x^k) + 2^{n-1}\log_2 k + O(1)$. Here, $P^k(x^k)$ is the maximum of the probability of x^k for order k Markov sources. Hence, disregarding the $O(1)$ term, the order k will yield a minimum codelength for x^k at

$$k^* = \arg\max \left[\log_2 P^k(x^k)\right] - 2^{n-1}\log_2 k .$$

This k^* is taken as the MDL estimate of the Markov order n.

This MDL has a strong consistency and maximizes the prediction efficiency of the model [12]. Our main aim is to apply the MDL-Markov analysis to the discussion process data in order to estimate the process structure with consistency and maximum prediction efficiency.

The Markov order in this analysis shows how many previous message a message in the discussion process depends on. When the Markov order of a discussion shows a large value, it means that the discussion process is colorful and fruitful. Conversely, when it shows a small value, it means that the discussion process is monotonous and simple.

4. Online MDL-Markov analysis system

We have developed an online MDL-Markov analysis system based on the ideas in the previous sections. This system estimates the Markov order n and the structure from the discussion processes' data based on the method described in section 2.

The estimated structure is visualized in the system as shown in Figure 2. The system shows the following results online:

1) A visualization of the estimated Markov structure with the order n, which maximizes the MDL,
2) The top 10 observed category sequences of the estimated Markov order pattern, and
3) The estimated Markov process structure.

The system updates the estimation values whenever the learners log in. By using this system, the learners can monitor the actual states of their discussion and improve their learning community.

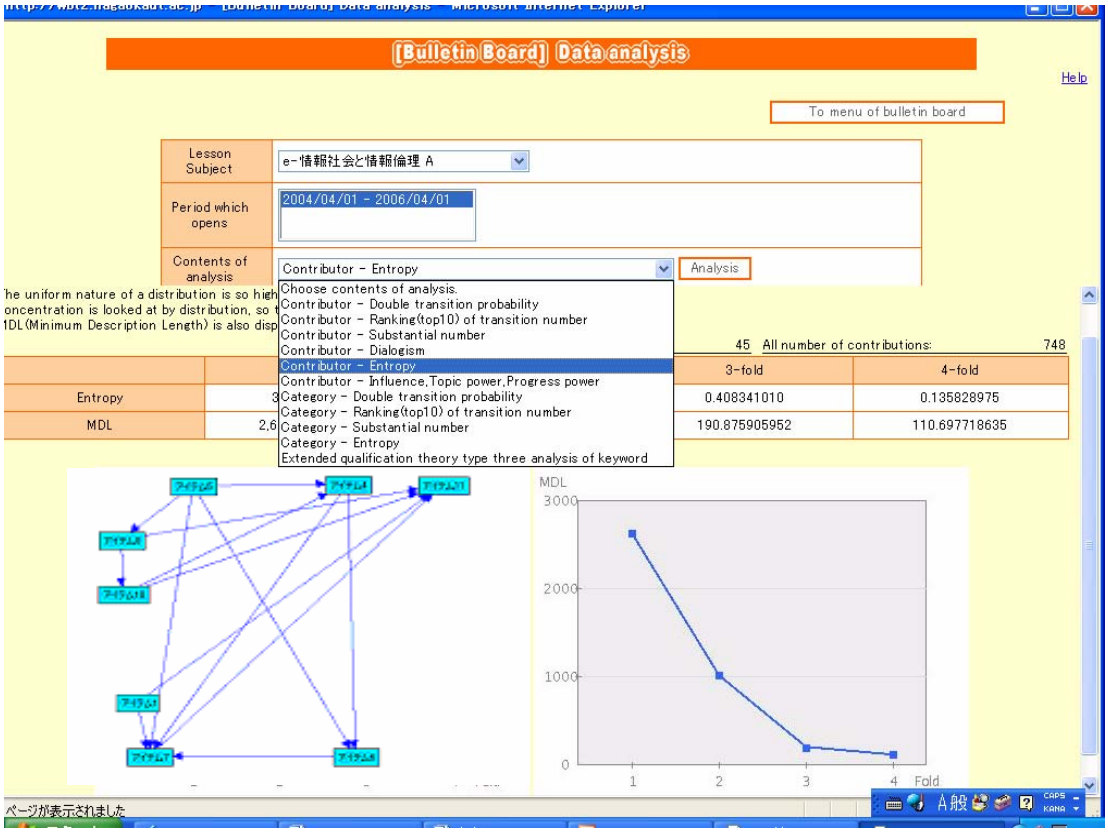

Figure 2. Online MDL-Markov analysis system

Table 1. Details of the analytical data

	Data 1	Data 2	Data 3
Subject name	"Information Society"	"Information Society"	"Safety and Management"
Year	2004	2005	2005
Number of learners	48	21	7
Number of contributions	748	217	21
Estimated Markov order	4	4	2
The top 5 observed category sequences of the estimated Markov order pattern. ( ) indicates the number of observations.	1, 1, 1, 1 (44) 5, 1, 1, 1 (25) 1, 1, 1, 9 (14) 1, 1, 1, 14 (7) 10, 1, 10, 9 (3)	2, 14, 5, 2 (38) 2, 2, 2, 2 (14) 2, 6, 2, 1 (13) 5, 2, 3, 2 (6) 4, 2, 2, 9 (4)	9, 14 (6) 2, 14 (4) 3, 14 (2) 8, 14(2) 10, 14 (1)

5. Examples of discussion processes analysis

This section demonstrates some examples of the online analysis results actually presented in the system for learners to monitor the collaborative learning states. Table 1 shows the details of three examples of analytical data from the system.

5.1. Data 1

Data 1 is the analysis data of the subject "Information Society" in 2005. Since the estimated Markov order is 4, this discussion process was comparatively colorful and fruitful. Furthermore, the Markov graph as shown in Figure 3 and the top 5 observed category sequences show that this discussion tended to center on category 1 (an opinion concerning the lecture content).

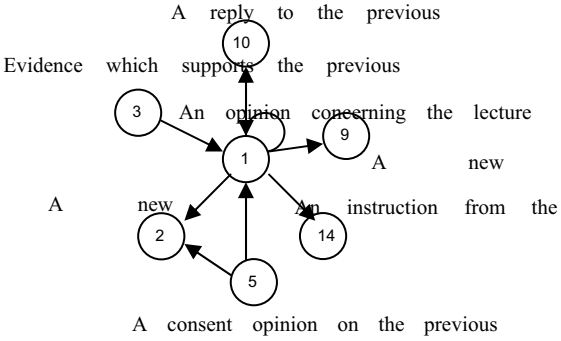

Figure 3. The estimated Markov graph for data 1

5.2. Data 2

Data 2 is the analysis data of the subject "Information Society" in 2005. That is, this is the same class discussed in data 1. The number of learners is 21. The analysis results shown here were presented in the system when the number of contributions was 217. From this data, the Markov order is estimated to be 4. This means that this discussion process was also comparatively colorful and fruitful. The Markov graph as shown in Figure 4 was estimated. Furthermore, the sequences "2, 14, 5, 2", "2, 2, 2, 2", "2, 6, 2, 1", "5, 2, 3, 2", and "4, 2, 2, 9" are the top 5 observed category sequences. These mean that the teacher was working well and question-answer interactions often appeared. This discussion tended to center on category 2 (a new opinion). Consequently, there are many new opinions and many interactions.

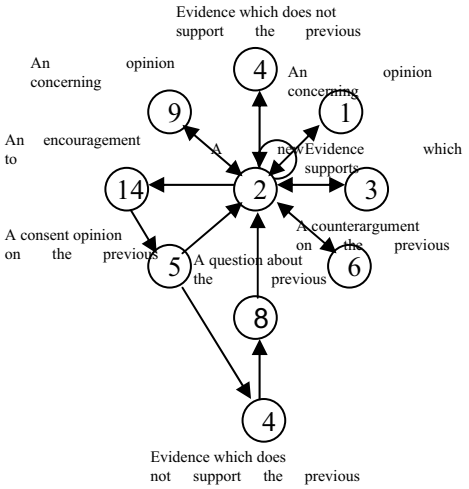

Figure 4. The estimated Markov graph for data 2

5.3. Data 3

Data 3 is analysis data of the subject "Safety and Management" in 2005. That is, this is the same class as data 1. The number of learners is 7. The analysis results shown here were presented in the system when the number of contributions was 21.

From this data, the Markov order is estimated at 2. This means that the discussion process was monotonous and simple at the time the data was presented. Furthermore, the Markov graph as shown in

Figure 5 and the sequences "9, 14", "2, 14", "3, 14", "8, 14", and "10, 14" as the top 5 observed category sequences interrupted the development of the discussion. This result reveals that the teacher must improve his methods in this discussion.

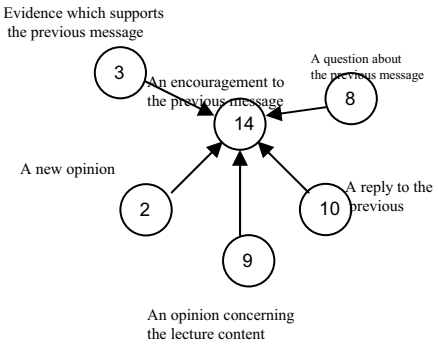

Figure 5. The estimated Markov graph for data 3

6. Evaluation

The proposed system was evaluated through interviews some of the learners and the teachers. The results from the interviews can be summarized as follows:

- The system visualizes the activities of a discussion process. Thereby, it enhances the motivation to make the discussion more fruitful.
- The system visualizes what is insufficient in a discussion. That is, the system derives some reflections of the current discussion. It shows the learners and the facilitators what they should do.

Thus, the results show that the proposed system motivated.

In addition, in questionnaires of the learners, 86% answered that this system was effective in improving their collaborative learning.

7. Conclusions

We have proposed an online visualization system of discussion processes in an electronic forum to assist learners by enabling them to monitor the actual states of their discussion thereby improving their learning community. The proposed method treats each message in the electronic forum as a code and assumes that a discussion process, which is defined by the data sequence of message categories used, follows a Markov source from an information theory approach. The proposed system estimates the structure of the Markov source using the MDL. from the discussion processes' data and visualizes the structure online. The advantages of the method are 1) it provides optimal data compression and 2) it has a strong consistency, which maximizes the prediction efficiency. The results of our evaluations of this system showed it to be effective.

8. References

[1]R. Palloff and K. Pratt, "Building Learning Communities in Cyberspace: Effective Strategies for the Online Classroom", Jossey-Bass, Inc. Pub., San Francisc, 1999.

[2]C. Gutwin, G. Stark, and S. Greenberg, "Support for workspace awareness in educational groupware", *Proceedings of CSCL'95 (Bloomington, IN, October 1995)*, LEA, 1995, 147–156.

[3]Y. Kimura and T. Tsuzuki, "Group Decision Making and Communication Model: An Experimental Social Psychological Examination of the Differences between the Computer-mediated Communication and the Face-to-face Communication", *The Journal of Experimental Social Psychology*, **38**, 2, 1998, 183–192.

[4] J.Nakahara,, S.Hisamatsu, Y.Yaegashi,, and Y.Yamauchi, "iTree: Does the mobile phone encourage learners to be more involved in collaborative learning? ", *Proceedings of Computer Supported Collaborative Learning 2005* ,Taipei, ILS, 2005, 198-242

[5]A. Martínez, Y. Dimitriadis, B. Rubia, E. Gómez, and P. de la Fuente, "Combining Qualitative Evaluation and Social Network Analysis for the Study of Classroom Social Interactions", *Computers & Education*, **41**, 4, 2003, 353–368.

[6]S. Puntambekara and R. Luckin, "Documenting Collaborative Learning: What Should be Measured and How?", *Computers & Education*, **41**, 4, 2003, 309–311.

[7]T. Mochizuki, S. Fujitani, Y. Isshiki, Y. Yamauchi, and H. Kato, "Assessment of Collaborative Learning for Students: Making the State of Discussion Visible for their Reflection by Text Mining of Electronic Forums", *Proceedings of E-Learn 2003* (Phoenix, AZ, November 2003), AACE, 2003, 285–288.

[8] T. Mochizuki, H. Kato, S. Hisamatsu, K. Yaegashi, S. Fujitani, T. Nagata, J. Nakahara, T. Nishimori, and M. Suzuki–449.

[9]M. Ueno, "Data mining and text mining technologies for collaborative learning in LMS 'SAMURAI'", Special Panel "Collaborative Technology and New e-Pedagogy", *Proc. of ICALT2004*, 2004, 1052–1053

[10]C.E.Shannon and W.Weaver, "*The Mathematical Theory of Communication*", Illini Books edition, 1963.

[11]J. Rissanen, "Modeling by Shortest Data Description", *Automatica*, Vol. 14., 1978, 1080–1100.

[12] J. Rissanen, "A universal prior for integrals and estimation by minimum description length", *Annals of Statistics*, **11**, 1983, 416-431.