System for Peer Review by Relative Evaluation in Group Learning

Takayuki Watabe
Graduate School of Informatics, Shizuoka University, gs11055@s.inf.shizuoka.ac.jp

Yoshinori Miyazaki
Faculty of Informatics, Shizuoka University, yoshi@inf.shizuoka.ac.jp

Abstract

We propose a method to evaluate individual learners in group learning. This method adopts a peer review system based on one-to-one comparison (relative evaluation) by the analytic hierarchy process (AHP). Moreover, in connection with incorporating relative evaluation, we believe it is necessary to consider how reliably each learner is capable of evaluating other members. Our proposal includes two algorithms for estimating the reliability. A pilot experiment for on-site learners was conducted. The results of analyses are shown and discussed.

1. Introduction

It is not easy to provide learners with fair evaluations in a setting for group learning. Conventionally, the same scores are given to all members of a group. For instance, the presence of an excellent learner leads to the provision of high scores to all, which is not necessarily a proper evaluation. This suggests a need for evaluating individual learners in a group. Therefore, we set a goal to propose a method for a fair evaluation by adopting a peer review system among group learners.

In general, evaluation criteria are divided into two categories: some require absolute evaluations and others relative evaluations. What we tackle in this paper is devising a method for relative evaluations. In a setting for group learning, the individual learners are enabled to make one-to-one comparisons between group members. As an algorithm for evaluation by one-to-one comparison, the analytic hierarchy process (AHP) [1] is incorporated. Although [2,3] attempted to apply AHP for peer review, our aim is further to make an appropriate ordering of learners by considering reliability of each learner's evaluation.

The proposed method was carried out by letting learners make relative evaluations in group work. As a result of the data analysis, some findings were obtained regarding relative evaluations.

Section 2 introduces some previous studies regarding peer review among learners. Section 3 elaborates on the proposed system configuration. Section 4 shows and discusses the results of data analyses of a pilot experiment. Section 5 presents some concluding remarks.

2. Literature review

As a review of computer-supported collaborative learning (CSCL), [4] investigates relevant papers for the present trends in this research field. This indicates that few cases have practical use, even though much research on CSCL has been conducted. It also indicates that the systems developed with file-sharing functions have low traceability regarding when, how, and by whom changes were made to which files.

Several pieces of literature related to peer evaluation (or mutual evaluation) are discussed below. [5] confirmed the higher maturity of understanding by learners in experiments conducted by assuming that the following five conditions lead to better collaborative learning: 1) giving a chance for peer evaluation; 2) keeping groups small, with at most five per group; 3) monitoring the activities of group members at random; 4) allocating roles to each member; and 5) having learners study collaborative learning beforehand. [6] obtained the result that an additional function of peer evaluation in collaborative learning can be to trigger a greater motivation to study. Conversely, [7] warns that as a con of peer evaluation there will be cases where the average point will depend on the evaluators if the learners are evaluated from subjective feelings. In comparison, our method can prevent such a problem by normalizing evaluations made by each learner. [8, 9] are pieces of research intended to gain high learning effectiveness through peer review. [8] had learners comment on writing assignments submitted by other learners, similar to an academic society in which a researcher submits a paper to a journal and receives reviews from referees. The learners are evaluated by a teacher based on the peer review. In [9], an attempt was introduced that learners revise their sentences after the process of peer reviewing with leaving comments in EFL (English as a Foreign Language) writing class.

CollabTech 2012 , August 27-29, 2012, Hokkaido, Japan.

[2] is similar to our present work as the AHP is incorporated in order to evaluate individual learners. What differs from ours is to make evaluations by the comparison between oneself and other learners. The reliability of evaluations is not counted. In [3], a Moodle module enabling teachers to analyze the study logs of learners with the AHP was developed, although the AHP was applied only to determine relative importance among the criteria used to evaluate learners, and not to evaluate themselves.

3. Outline of the proposed method

Our method is thoroughly explained in this section. First, how learners evaluate other learners is given. Next, two algorithms to combine each learner's evaluation are shown. The difference between the two algorithms comes from the different perspectives on the reliability of evaluation.

For the rest of this paper, assume that each group consists of four members denoted by $L_i (i = 1 \text{ to } 4)$.

3.1. Scoring algorithm for peer review system

The peer review algorithm comes first. A learner makes one-to-one comparisons between any two other learners in the group to which the learner belongs. Learners do not compare with themselves. This is to prevent learners from intentionally evaluating themselves highly to raise their own scores. Therefore, each learner makes three comparisons for every criterion; for example, L_1 compares L_2 with L_3, L_3 with L_4, and L_4 with L_2. Of course, the more group members there are, the more times learners have to compare. In fact, letting the number of group members be n, the number of comparisons becomes ${}_{n-1}\mathrm{C}_2 = (n-1)(n-2)/2$ in general. On the other hand, [5] reports that a group learning be conducted at a party of at most five learners, as stated above. This implies that the additional burden by one-to-one comparisons will not be a major issue.

In comparing two learners, nine scores are possible, depending on the degree of difference between the two. A maximum of 4 is scored if L_i is far superior to L_j. Conversely, a minimum of –4 is scored if L_i is far inferior to L_j. If L_i and L_j are equal, the score is zero. Thus, the intermediate values (3 through 1 and –1 through –3) reflect the extent of the difference. Let $E(L_i, L_j)$ denote the evaluation of L_i against L_j. Namely, $E(L_i, L_j) = 4$ indicates that L_i is far superior to L_j. For example, suppose that L_1 evaluates the others as $E(L_2, L_3) = 2$. It follows that $E(L_3, L_2) = -2$.

3.2. Computational algorithm for importance of learners

Next, an algorithm is provided for ordering learners as a result of the peer review. This algorithm is basically from the AHP.

Although the range is centered on zero to allow intuitive evaluations by learners, each original value a's must be processed into a value c's in such a way that the AHP may be applied:

1. $b \leftarrow \mathrm{sign}(a) \cdot (|a| + 1)$,
2. $c \leftarrow 1/|b|$ if b is negative, $c \leftarrow b$ otherwise,

where $\mathrm{sign}(x)$ is a function that returns 1 if $x \geq 0$ and –1 otherwise. Consequently, the range is changed from {–4, –3, –2, –1, 0, 1, 2, 3, 4} to {1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5}.

Let accordingly processed values be inserted into a table so that the values of $E(L_i, L_j)$ are placed at the intersection of the L_i-row with the L_j-column (Figure 1). Furthermore, let this table be represented in the form of a matrix.

	L_2	L_3	L_4
L_2	1	5	3
L_3	1/5	1	1/5
L_4	1/3	5	1

$$\Rightarrow \begin{pmatrix} 1 & 5 & 3 \\ 1/5 & 1 & 1/5 \\ 1/3 & 5 & 1 \end{pmatrix}$$

Figure 1. Processed evaluations by L_1.

Let us refer to this matrix as a pairwise comparison matrix. What remains is to compute an eigenvector corresponding to the maximum eigenvalue of the matrix. An eigenvector of this example is $(0.617, 0.086, 0.297)$ and each element of this eigenvector suggests the relative importance of a learner (L_2, L_3, or L_4). In this example, L_1 regards L_2 and L_3 as the most and least excellent, respectively.

3.3. Measuring for the reliability of evaluations

As the last section suggests, four eigenvectors representing relative evaluations by four learners are obtained for every criterion. However, to order the learners, combining these multiple evaluations is necessary. In this work, we propose two algorithms for the measurement of qualities of evaluations, and a method to combine the computed evaluations.

3.3.1. Autoregression model. This section deals with a measurement of reliability, based on the assumption that excellent learners can produce reliable evaluations.

This algorithm is analogous to PageRank [10], an algorithm used for link analysis of web pages.

Let us give an example. Suppose that as a result of relative evaluations the learners of a group obtained the following eigenvectors (Figure 2).

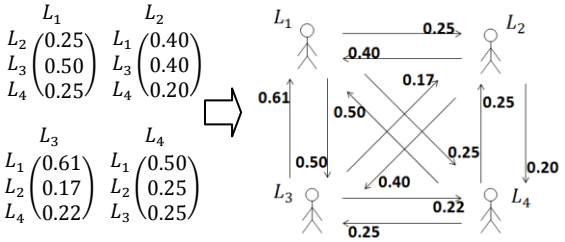

Figure 2. An example of eigenvectors and the graph representation.

Let the initial value be 0.25 for each learner. What L_1 receives from L_2 is (the value from L_2's value $\times$ the importance given to L_1 by L_2), which is 0.25×0.40. Likewise, L_1 receives 0.25×0.61 from L_3, and 0.25×0.50 from L_4. The next step is to update the value for L_1 by adding the three contributions. The values for L_2, L_3, and L_4's are updated in the same way. The final step is to normalize the values for L_1, L_2, L_3, and L_4 such that the sum equals unity. The whole process is continued until the values for L_1, L_2, L_3, and L_4 each converge. The converged values are the scores of the learners. In this example the values for L_1, L_2, L_3, and L_4 are 0.343, 0.180, 0.290, and 0.187, respectively.

3.3.2. Application of C.I. Another measurement of reliability is based on the premise that consistent evaluations are reliable. A consistency of evaluations is explained with an example as follows. Assume that the evaluation by L_1 went $L_2 > L_3$ and $L_2 < L_4$. Thus, it should follow that $L_3 < L_4$, considering the two inequalities. On the contrary, if another evaluation went $L_3 > L_4$, it may be remarked that the evaluation by L_1 lacks consistency. If the evaluation by some learner is considered inconsistent, the evaluation is ignored, judging the evaluation unreliable.

An index called C.I. (Consistency Index) may be used for the consistency of evaluations. The C.I. is computed as

$$\text{C.I.} = \frac{\lambda_{\max} - n}{n - 1},$$

where $\lambda_{\max}$ denotes the maximum eigenvalue of the given pairwise comparison matrix and n is the matrix size. When the value of the C.I. exceeds 0.1, the consistency of the evaluations is suspect. In the above case $\text{C.I.} = 1.09$, is obtained.

For instance, suppose that the C.I. for L_1 is greater than 0.1 and the following eigenvectors are obtained from the evaluations by L_2, L_3, and L_4 (Figure 3):

$$L_2 : \begin{matrix} L_1 \\ L_3 \\ L_4 \end{matrix}\begin{pmatrix} 0.41 \\ 0.26 \\ 0.33 \end{pmatrix}, \quad L_3 : \begin{matrix} L_1 \\ L_2 \\ L_4 \end{matrix}\begin{pmatrix} 0.10 \\ 0.45 \\ 0.45 \end{pmatrix}, \quad L_4 : \begin{matrix} L_1 \\ L_2 \\ L_3 \end{matrix}\begin{pmatrix} 0.49 \\ 0.31 \\ 0.20 \end{pmatrix}$$

Figure 3. Example of a set of eigenvectors less that of L_1.

Accordingly, the score of L_1 is set to the average of the scores given to L_1 by L_2, L_3, and L_4 toward . Or, which is $(0.41 + 0.10 + 0.49)/3$. The score of L_2's is set to the average of the scores given to L_2 by L_3 and L_4, which is $(0.45 + 0.31)/2$. Likewise, the scores of L_3 and L_4, are obtained by taking the averages of the scores by the remaining two learners other than L_1.

Note that this method cannot be applied if there are three or more learners out of four whose C.I. values exceed 0.1. In that event, all the learners are given the same value ($= 0.25$), indicating that the peer review within the group does not make any sense.

3.4. Weighting algorithm for each criterion

Our method also combines learners' scores based on different criteria to produce their final scores. This weighting procedure is performed by teachers, not by learners. One-to-one comparisons are again used to weight each criterion.

4. Experiments and discussions

A pilot experiment was conducted to determine the effects and validity of the method proposed in section 3. In the last three classes of the "discrete mathematics" course held at the Faculty of Informatics, Shizuoka University, in 2011, six groups, each consisting of four members, were formed to perform group work. Criteria were set as the "amount of effort the learner devoted to his/her allocated task (c1)," "quality of the task given to the learner (c2)," and "contribution to group work by communicating with other members (c3)."

4.1. Discussion on the variety of scores by evaluators

As a result of peer reviewing, a large number of values of the one-to-one comparisons were zero. More precisely, 69% of the total evaluations were zero. The presence of many zeros follows the difficulty in clarifying differences among learners, leading to the peer review system having little significance.

The cause of this phenomenon is assumed to be an inappropriate setting that allows evaluators to choose

values from the range. It may not be difficult to determine which of two learners is superior, whereas determining this with the inclusion of the extent may be burdensome. In order to reduce this burden, it may be necessary to decrease the number of choices to three, or, "$L_i > L_j$," "$L_i = L_j$," and "$L_i < L_j$." More extremely, one may eliminate the option "$L_i = L_j$."

4.2. Evaluation of the values obtained by the algorithms

We discuss the results of the method based on the autoregression model (referred to as Method 1) and another method taking advantage of C.I. (Method 2).

For each criterion, learner ranking was attempted using the total sum of corresponding eigenvector elements obtained from the three classes. Weighting algorithms among the criteria were ignored in this attempt. Rankings among four learners were obtained for six groups and three criteria. This was followed by computing rank correlations among the thus obtained rankings and those from the final exam. The averages of six groups for each criterion are shown in Table 1.

Table 1. Rank correlation coefficients between rankings from Methods 1&2 and the final exam

Criterion	Method 1	Method 2
c1	-0.148	-0.315
c2	-0.148	0.119
c3	0.352	0.319

As shown by the results, the evaluation method adopted in this work did not lead to the acquisition of strong correlations. For c1 in particular, the results showed negative correlations in both methods. On the other hand, a certain degree of positive correlation was admitted for c3.

The observed tendency is that the difference between the two methods is comparatively larger when the number of $E(L_i, L_j) = 0$ is small. It is expected that the difference between the two algorithms will become significant upon improvement of the relative evaluation method mentioned in section 4.1 because a larger set of evaluations will be considered, in which the number of $E(L_i, L_j) = 0$ is small.

5. Concluding remarks

In this work, methods were proposed for realizing a relative evaluation in the form of peer reviewing in order to rank learners in a group.

Our future work consists of two parts. The first is the improvement of the multiple choices presented to learners when peer reviewing is performed. From the experiment, it was revealed that learners were likely to judge one-to-one comparisons as equivalent. The frequency of such judgments has to be reduced, especially for the demand to rank learners.

The second is to devise a method to evaluate the results of the peer reviewing process. As mentioned in section 4.2, this paper made an attempt to evaluate the correlation with final exam scores. An alternative method is comparing peer review results with learners' activities on an LMS (Learning Management System), such as log data on the number of times learners refer to study materials pages and solve assignments.

References

[1] T.L. Saaty, *The Analytic Hierarchy Process*, McGrawHill, New York, 1980.

[2] Y. Ueda, "An Attempt of Peer Review in Small-sized Class by Analytic Hierarchy Process (AHP)", *International Buddhist University Review*, Vol. 41, 2006, pp. 159-169. (in Japanese)

[3] Y. Miyazaki, and Y. Sugiura, "Module Development for Learners' Study Logs and Peer Evaluations for CSCL", *CollabTech2008*, 2008, pp.84-88.

[4] T. Kojiri et al., "Computer Supported Collaborative Learning (CSCL) and Support Technology", *Transactions of Japanese Society for Information and Systems in Education*, Vol.23, No.4, 2006, pp. 209-221. (in Japanese)

[5] S. Ochoa et al., "Improving Learning by Collaborative Testing", *Journal of Student-Centered Learning*, Vol.1, No.3, 2003, pp. 123-135.

[6] T. Akakura, and G. Nana, "A Development and an Evaluation of the E-Learning System to Promote Group Learning by the Function of Mutual Evaluation in Group", *Forum on Information Technology*, 2007, pp. 525-526. (in Japanese)

[7] Y. Fujiwara et al., "Evaluator Selection Algorithm for Mutual Evaluation in a Learning Community", *The Institute of Electronics Information and Communication Engineers Technical Report of IEICE.ET*, Vol.104, No.703, 2005, pp. 97-100. (in Japanese)

[8] E.Z.F. Liu et al., "Web-Based Peer Review: The Learner as both Adapter and Reviewer", *IEEE Transactions on Education*, Vol. 44, No. 3, 2001, pp. 246-251.

[9] H.C. Liou, and Z.Y. Peng, "Training Effects on Computer-mediated Peer Review", *System*, Vol. 37, No. 3, 2009, pp. 514-525.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", *Technical Report*, Stanford InfoLab, 1999.