A Students' Mutual Evaluation Method for Online Reports using Groupware

Yuanyuan WANG†, Ryosuke YAMASHITA†, Yukiko KAWAI†, Setsuko MIYAMOTO††, and Kazutoshi SUMIYA††

† Faculty of Computer Science and Engineering, Kyoto Sangyo University
Motoyama, Kamigamo, Kita-ku, Kyoto 603-8555, Japan
†† School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †{yuanw,kawai}@cc.kyoto-su.ac.jp, ††g1145424@cse.kyoto-su.ac.jp,
†††{miyamoto,sumiya}@shse.u-hyogo.ac.jp

Abstract Nowadays, mutual evaluation in education is an available method that students in the learning community use to evaluate each other. The method calculates scores of students' reports by considering who evaluates those reports. In this paper, we propose a student's mutual evaluation method using the PageRank algorithm, an appropriate evaluation method that helps teachers to easily understand whose report is the best from the students' viewpoints. In particular, we perform students' mutual evaluation based on a groupware by utilizing a "Like" function in a course practice. As a result, it was able to not only provide an overall rating from the students' sum of votes but also, by considering who voted, to promote the reliability of the students as evaluators for their mutual evaluation.

Key words students' mutual evaluation, groupware, online reports, voting, posting time

1. Introduction

Web-based report systems, such as Bulletin Board Systems (BBSs) and groupware, are now one of the most frequently used tools in e-learning currently. Students then post and share their reports at anytime and from anywhere in a given period, i.e., after a lecture and before the next lecture, and the students can browse and vote other students' reports through these online systems. However, teachers need to evaluate all students' online reports, but this will require a great deal of time and effort for a fair and multi-faceted evaluation of the reports.

As depicted in Figure 1, we propose a students' mutual evaluation method to enable students instead of teachers to evaluate their reports by voting with each other. It provides scores of reports by analyzing the relationship between voting and posting time of the reports based on a voting graph of the reports, to promote the quality of the votes and prevent unfair votes. For this, the voting graph is constructed by the votes between a student and his or her voted reports. In this paper, we perform a students' mutual evaluation using groupware based on voting with a "Like" button in a course practice. Students can vote on the others' reports by pressing a "Like" button, when they think the report is

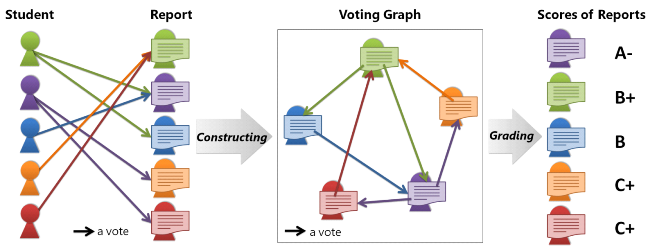

Figure 1 Conceptual diagram of our students' mutual evaluation method

good [3]. It can reduce the students' burden of evaluating others' reports without specific points. With our students' mutual evaluation method, teachers can efficiently acquire a score ranking list of students' reports through groupware. Moreover, students can easily detect best reports from the score ranking list of their reports.

The next section describes our proposed report scoring method based on students' mutual evaluation. In Section 3, we summarize the results of the proposed method in a course practice. Finally, in Section 4, we conclude this paper with suggestions for future work.

Figure 2 Steps of a course using groupware system

2. Students' Mutual Evaluation Method using Groupware

During a course using the groupware system (Cybozulive(注1)) is shown in Figure 2, the following steps are followed: 1) posting reports after lectures in a certain period, 2) browsing reports written by other students and voting with a "Like" button; and 3) receiving votes for their own reports.

2.1 Construction of Voting Graph and Transition Probability Matrix

A voting graph is first constructed. The nodes of the directed graph consist of students' reports, and the links can be considered as votes from students for the others' reports. In our previous work, we built a system that evaluates users who browse the Web pages based on their links between a user and his or her browsing pages [5]. In this work, in order to evaluate the reports based on students' mutual evaluation; we focused on the students who vote on the others' reports. If one student u_i is voting another student u_j's report, then, a link from u_i's report to u_j's report (arrows in Figure 1(1)(2)), and the element of its corresponding adjacency matrix (u_j, u_i) is set to 1. As an example shown in Figure 1(3), the elements (u_1, u_4), (u_1, u_5), (u_2, u_1), (u_2, u_3), (u_3, u_1), (u_4, u_2), (u_5, u_2) become 1.

We next describe the construction of the transition probability matrix from the adjacency matrix. For example, it can be transformed to a transition probability matrix as shown in Figure 1(4). Suppose that the students often post their reports before by referring to previous others' reports, we should reduce the scores of the last posting reports. Then, w_i is the weight assigned to each student's report by considering its posting time.

2.2 Score Calculation for Reports

We next calculate the scores of students' reports by the following formula based on the concept of PageRank [2] and ObjectRank [1].

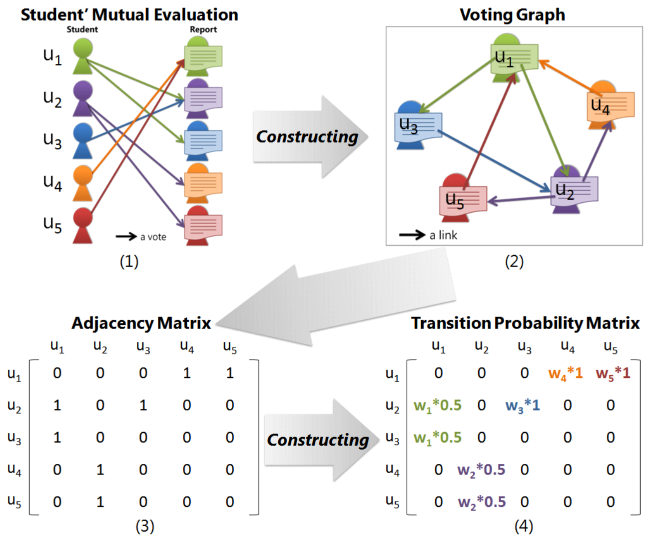

Figure 3 Voting graph and transition probability matrix

$$S(r) = (1-d) + d * (\frac{S(v_1)}{T(v_1)} * w_1 + \cdots + \frac{S(v_i)}{T(v_i)} * w_i) \quad (1)$$

- r: a student's report, i.e., the report of u_1, u_2, u_3, u_4, or u_5
- $v_1, \cdots, v_i$: the set of the votes of r, i.e., the votes of u_1's report by u_4 and u_5
- $S(v_i)$: the numerical weight of each vote contained in the set of the votes of a student's report
- $T(v_i)$: the number of votes from a student, i.e., $T(v_1)$=2
- w_i: the weight of posting time of a report, i.e., w_1 is the weight of u_1's report. There are two methods:
  (1) w_i=$1/n_i$, n_i is the posting number of u_i's report
  (2) w_i=$(m$-n_i+$1)/m$, m is the total number of students in a course
- d: a damping factor adjusts the derived value downward. Various studies have tested different damping factors, but it is generally assumed that the damping factor is set at approximately 0.85

Initially, the weight of each vote is 1, if a student votes multiple report, the weight distributes through each vote evenly, e.g., if u_1 votes reports of u_2 and u_3, the weight of each vote becomes 0.5. Finally, we normalize the score of each report

(注1)：https://cybozulive.com/

between 0–10.0.

Using our proposed mutual evaluation method, 1) a report of a student who votes for other students' reports produces a better report himself or herself and 2) a report with many votes raises the reliability of its author's opinion when he or she casts his or her own votes. Thus, teachers and students can easily understand whose report is the best from the students' viewpoints, and we believe that this method can lead to an appropriate evaluation method in education in the future.

3. Evaluation

3.1 Prototype System

Based on the method described above, we have built a system to support report scoring, using Python 2.7.8. The interface is programmed using Tkinter (GUI: graphical user interface). The prototype system has two stages: analysis and calculation. Firstly, in the analysis stage by using our developed Vote Checker (left part of Figure 4), we first construct a directed graph of votes, which consists of students' reports as the nodes by analyzing how many votes are received of each report and who voted. Then, its corresponding adjacency matrix is constructed. Secondly, in the calculation stage by using our developed Score Calculator (left part of Figure 4), the adjacency matrix is transformed to a transition probability matrix with the weight of the posting time of the reports, then, the scores of the reports are calculated based on the transition probability matrix, and the scores are ranked in an order from high to low.

3.2 Results of Students' Mutual Evaluation

In this section, we present our findings from the results of our proposed report scoring method based on students' mutual evaluation in a course practice. This is a course of Applied Informatics, which consists of 10 lectures on different topics, and 20 students who participated in this course. Using the "Like" button as a vote through an online groupware, (1) the students must to browse any other's report (need not to browse all others' reports) and vote on it, when they think it is good; (2) each student votes at least one report and up to five reports. We calculated the scores of the reports as follows:

- A) Baseline 1: counting the sum of the number of "Like" of a reports from other students
- B) Baseline 2: calculating TF-IDF values of a reports by considering the content analysis
- C) Baseline 3 [6]: counting the number of words in a report by considering the size of the report
- D) Previous [7]: using Eq. (1) without the weight w_i by considering the quality of votes only
- E) Proposed 1: using Eq. (1) with the weight $w_i(=1/n_i)$ by considering both the quality of votes and the posting time
- F) Proposed 2: using Eq. (1) with the weight $w_i(=(m-n_i+1)/m)$ by considering both the quality of votes and the posting time

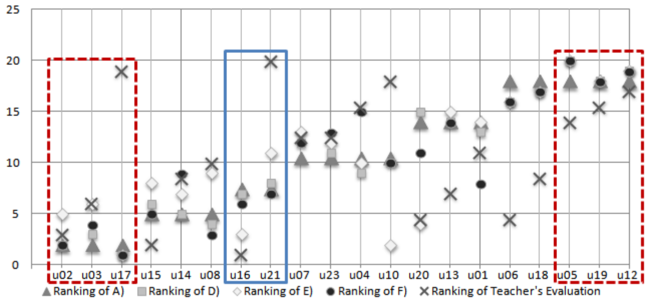

Figure 5 Correlation diagram

Figure 2 shows a correlation diagram of rankings of the methods, A), D), E), and F) by using the "Like" button and the teacher's evaluation of Lecture #3, the horizontal axis denotes student numbers of the reports in an order of the ranking based on the method A), and the vertical axis denotes the ranking number. Here, the teacher emphasized on the content of the report to evaluate them. The results and our findings were as follows:

- Although, some reports gained the same number of votes by baseline A); their scores were different by our previous method D), and our proposed methods, E) and F). For example, the scores for the reports of students, u16 and u21, were identical by A) (frame in Figure 2). However, they were different based on D), E), and F).
- The top ranked reports have high scores by A), D), E), and F), and they were correlated with the teacher's evaluation. For example, the scores of the top ranked three reports of students, u02, u03, and u17, were all high by A), D), E), or F) (left dashed frame in Figure 2). Except u17's report, the scores of the reports of u02 and u03 were all high by the teacher's evaluation.
- The lowest ranked reports have low scores by all above methods, and they were correlated with the teacher's evaluation. For example, the scores of the lowest ranked three reports of students, u05, u19, and u12, were all low by A), D), E), or F) (right dashed frame in Figure 2), and they were also low by the teacher's evaluation.

For each report scoring method, we then calculated the Spearman's rank correlation coefficient [4] between the score rankings by the teacher's evaluation and the all above methods of Lecture #3. The correlation value ranges from -1 to 1, where -1 indicates that two rankings are completely reversed, whereas 1 indicates that the rankings are exactly the same. The correlation results are listed in Table 1, and the

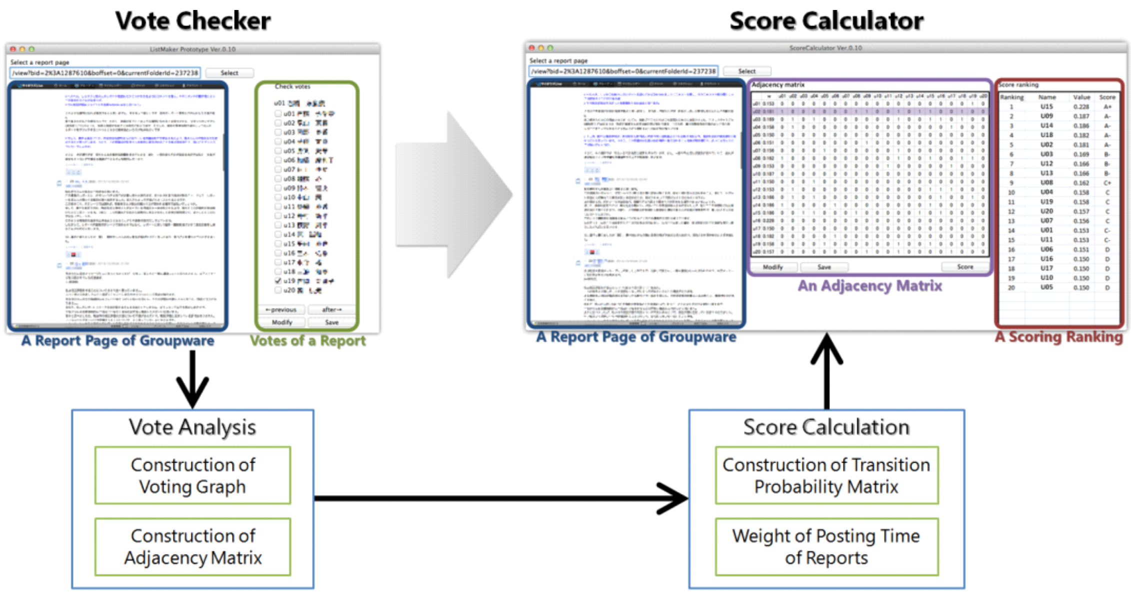

Figure 4 Screenshot of our prototype system

Table 1 Correlation results

Report Scoring Method	Correlation Value
A) Baseline 1	0.464
B) Baseline 2	0.458
C) Baseline 3	**0.583**
D) Previous	0.456
E) Proposed 1	**0.475**
F) Proposed 2	**0.507**

Table 2 Correlation results of combined methods

Combined Method	Correlation Value
D1) Previous×Baseline 3	0.522
E1) Proposed 1×Baseline 3	**0.593**
F1) Proposed 2×Baseline 3	**0.615**

results can be explained as follows:

- The correlation values of all above methods and the teacher's evaluation were not very close to 1.
- The correlation value of the baseline C) and the teacher's evaluation was highest than other methods, A), B), D), E), and F).
- The correlation values of our proposed methods, E) and F), and the teacher's evaluation were higher than those of the methods, A) and D) based on the "Like" button.

Since our proposed methods based on the "Like" button did not reach a very high correlation value and the size of the reports reached a highest correlation value, we combined the size of the reports and our previous work, proposed methods. We also calculated the Spearman's rank correlation coefficient between the score rankings by the teacher's evaluation and those combined methods of Lecture #3, and the correlation results are listed in Table 2

Although the correlation value of the combined method D1) was lower than the size of the reports, it was higher than D) and the combined methods, E1) and F1), were more higher than E) and F), this experiment indicated that our proposed report scoring methods combined with the size of the reports have the potential to support teachers easily and efficiently evaluate students' reports based on students' mutual evaluation with the "Like" button and the posting time without content analysis of the reports using groupware. Since our proposed methods by considering both the quality of the votes and the posting time of the reports (strategy aspect of utilizing the cultural psychology of Japanese) combined with the size of the reports, achieved a good performance compared with the conventional report scoring methods by counting the total number of votes or considering the content analysis (TF-IDF values).

Future work will deeply analyze the correlation between our proposed methods based on students' mutual evaluation and teachers' evaluation with large datasets. In the teachers' evaluation, we need to adopt different methods by considering the posting time of reports or not. In order to verify the reliability of our proposed methods, we should try to use other factors of the reports with the students' mutual evaluation.

4. Conclusions

In this paper, we proposed a report scoring method based on students' mutual evaluation using groupware. In a course practice, students performed a peer evaluation of their reports by voting for valuable reports using a "Like" button. Therefore, it is not only a total number of votes for evaluating the reports, but also considering both the relationship between voting and posting time of the reports. It can lead to a new method rooted in the indigenous culture of review by the students' mutual evaluation.

In the future, we need to measure inter-rater reliability of our proposed report scoring method by considering other conditions, e.g., voting time, similarity of reports, etc..

Acknowledgments

This work was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan, and JSPS KAKENHI Grant Number 26280042.

References

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proc. of VLDB 2004*, pages 564–575, 2004.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[3] J. Iturrioz, I. Azpeitia, and O. Díaz. Generalizing the like button: empowering websites with monitoring capabilities. In *Proc. of SAC 2014*, pages 743–750. ACM, 2014.

[4] J. L. Myers and A. D. Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, 2003.

[5] Y. Shiraishi, J. Zhang, Y. Kawai, and T. Akiyama. Simultaneous realization of page-centric communication and search. In *Proc. of CIKM 2012*, pages 2719–2721, 2012.

[6] M. Tsubaki, T. Oya, and T. Kobayashi. Analysis of the design of effective learning activities using learning–type pdca and capd cycles on the basis of the characteristics of individual students. *International Journal of Knowledge and Learning*, 8(1):150–165, 2012.

[7] Y. Wang, Y. Kawai, S. Miyamoto, and K. Sumiya. A students' mutual evaluation method for their reports using pagerank algorithm. In *Proc. of ICCE 2014*, pages 113–115, 2014.