

Crowdsourcing-Based Framework for Teaching Quality Evaluation and Feedback Using Linguistic 2-Tuple

Tiejun Wang¹, Tao Wu^{1,*}, Amir Homayoon Ashrafzadeh² and Jia He¹

Abstract: Crowdsourcing is widely used in various fields to collect goods and services from large participants. Evaluating teaching quality by collecting feedback from experts or students after class is not only delayed but also not accurate. In this paper, we present a crowdsourcing-based framework to evaluate teaching quality in the classroom using a weighted average operator to aggregate information from students' questionnaires described by linguistic 2-tuple terms. Then we define crowd grade based on similarity degree to distinguish contribution from different students and minimize the abnormal students' impact on the evaluation. The crowd grade would be updated at the end of each feedback so it can guarantee the evaluation accurately. Moreover, a simulated case is shown to illustrate how to apply this framework to assess teaching quality in the classroom. Finally, we developed a prototype and carried out some experiments on a series of real questionnaires and two sets of modified data. The results show that teachers can locate the weak points of teaching and furthermore to identify the abnormal students to improve the teaching quality. Meanwhile, our approach provides a strong tolerance for the abnormal student to make the evaluation more accurate.

Keywords: Teaching quality evaluation, crowdsourcing, linguistic 2-tuple, group decision making.

1 Introduction

Teaching quality is a key factor influencing student achievement, especially classroom teaching quality in university. The graduate attributes to professional engineering defined in Washington Accord [International Engineering Alliance (2014)] consists of 12 parts, and most of which obtained through studies in a classroom as class-based teaching is the main channel to impart knowledge and develop various capabilities to solve complex engineering problems. Therefore, it is more important to improve classroom teaching quality. Not only the teaching management department in universities but also the teachers themselves would like to know about the achievement of some knowledge points. Several evaluation mechanisms [Dong and Dai (2009); Chen, Hsieh and Do (2015); Chang and Wang (2016); Zhang, Wang, Zhang et al. (2017)] had been established to

¹ Department of Computer Science, Chengdu University of Information and Technology, Chengdu, 610025, China.

² CSIT Department, School of Science, RMIT University, Melbourne, 3058, Australia.

* Corresponding Author: Tao Wu. Email: wut@cuit.edu.cn.

leverage systematical and scientific methods to evaluate course quality and help teachers to improve the teaching methods. Normally, teaching quality evaluation can be regarded as a multi-attribute group decision making (MAGDM) problem, and similar problems had been discussed. In this issue, different kinds of questionnaires designed according to the teaching process including teaching preparation, teaching organization, teaching content, teaching methods, teaching effectiveness, etc. In the process of evaluating teaching quality, the assessment results obtained from the filled questionnaires. After that, teachers could get feedback about those key indicators from other people who could be students or other experts. The evaluation system getting feedback from experts could provide results about how a teacher or course is good enough to develop students' skills in a certain field as well as which part there are problems in the course. However, it could not give teachers the details information about which part of each course has problems as generally those experts have different backgrounds about responsible courses with the evaluated teachers. The other kind of evaluation system is getting feedback from students at the end of the semester as it also could not provide a specific solution since the feedback is seriously lagging behind.

Crowdsourcing is a sourcing paradigm in which individuals or organizations obtain goods and services from a large and relatively open group of internet users. It can divide complex problem among participants to achieve a cumulative result which could be ideas and finances. Some research [Luca and Michael (2014); Xu, Liu, Yen et al. (2017); Fernandes, Nati, Loumis et al. (2015)] have been done to collect information from a large crowd to help improve service quality. If teachers could get feedback about courses directly from a large number of students who are sitting the classroom, then that information without delay would be more accurate.

In this paper, we show a framework to aggregate the heterogeneous feedback directly from students seating in the classroom to generate more accurate evaluation conclusion by answering some simple questions about the courseware through the App or web browser. And when teachers go home, they could obtain the evaluation results to different knowledge points designed by themselves before class. As the assessments are for the courseware concerning some specific knowledge points that teachers care about, educators could utilize the evaluation results to refine the teaching content and method at low score points and finally make their course quality will increase respectively. To make the results more accurate, teachers could assign tasks only to specific students whose properties satisfied with instructor's requirements. Then the main issue is that how to provide an incentive for students to be accurate in their ranking. Furthermore, teachers can get details from different dimensions according to the student's properties.

The rest of the paper organized as follows: the related works will be in the next section, and in Section 3 we will show a framework based on crowdsourcing to evaluate teaching quality using linguistic 2-tuple model. Section 4 presents an example using our proposed framework. In the last section, we point out some conclusions and future works.

2 Related works

Crowdsourcing has been widely using in various fields to gather information and aid decision-making. Carrasco et al. [Carrasco, Sánchez-Fernández, Muñoz-Leiva et al.

(2017)] present a linguistic multi-criteria decision-making model for aggregating heterogeneous questionnaires with opinions about the quality of the e-service offered by the hotels. Authors gather information from several websites and find that all the scales slightly better evaluated. But it is too difficult to evaluate e-service by scales. Then Jatoth et al. [Jatoth, Gangadharan, Fiore et al. (2018)] addressed a hybrid multi-criteria decision-making model involving the selection of cloud services among the available alternatives by ranks. The proposed methodology assigns various ranks to cloud services based on the quantified quality-of-service parameters using a novel extended grey technique for an order of preference by similarity to ideal solution integrated with the hierarchical analytical process. Besides, CrowdGrader [Luca and Michael (2014); Xu, Liu, Yen et al. (2017)] is a system that allows students to submit and collaboratively review and grade homework. In CrowdGrader, students would benefit from being able to examine the solutions offered by other students. Meanwhile, all submissions will be distributed to different students to get some reviews. This system has been used at seven classes and utilizes grade to evaluate submissions. But the detail information of the ranks or grades could be lost in the evaluation.

Herrera et al. [Herrera and Martinez (2000)] found that the linguistic domain treated as continuous while in the symbolic model treated as discrete. They proposed a linguistic, computational model based on linguistic 2-tuples carries out processes of computing with words easily and without loss of information. After that, many research [Wang and Hao (2006); Merigo, Casanovas and Martinez (2010); Wei and Zhao (2012); Liu, Lin and Wu (2014); Wu, Wu, Zhou et al. (2015); Li, Dong, Herrera et al. (2017); Wei, Alsaadi, Hayat et al. (2018)] had been done from various aspects.

3 Proposed framework for teaching quality evaluation

Generally, digital courseware, such as PowerPoint and Keynote, is extensively used as the primary way to communicate with students in classrooms of universities. Digital courseware is the basement to evaluate teaching quality as they almost assume all the responsibilities to pass on knowledge to students. So we will show an approach, as shown in Fig. 1 that teachers can utilize to share courseware with students and mark some questions on specific slides to generate questionnaires at the same time. Then, students can make responses using applications installed in their smartphones or browsers in computers while the contents (knowledge points) in the marked slides taught by teachers in the classroom. Because the questions will only be displayed just when the related knowledge point illustrated, therefore, the feedback directly from students seating in the classroom or front of computers is more accurate. In this framework, the evaluation of teaching quality is launched by teachers, the feedback is coming from students, and the results are aggregated from that feedback and could be applied by teachers to refine the target content in this course. Thus, the problem reduced to aggregate relevant results according to the student evaluation of target knowledge points.

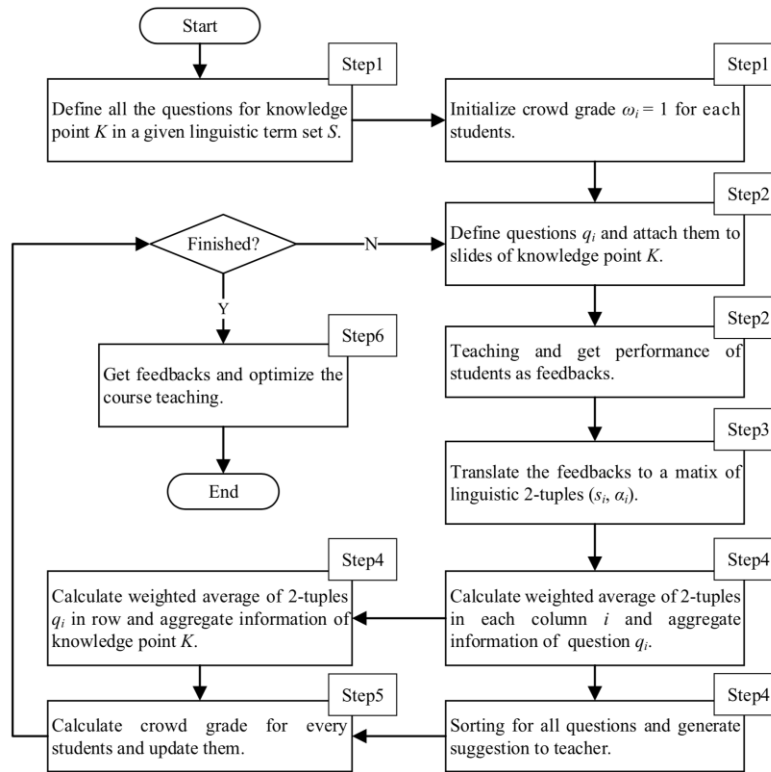


Figure 1: Flowchart of the proposed framework for teaching quality evaluation

3.1 Step 1 initiation

Usually, the syllabus for one course consists of some knowledge points that may reflect in different types of material, and teachers can organize them together to make digital courseware depending on some teaching methods. In the course design phase, a teacher may concern with the achievement of some knowledge point covered by some slides in digital courseware. After the slides associate with the knowledge point finished, teachers may ask some questions to students $X = \{x_1, x_2, \dots, x_H\}$ and hopefully get some feedback to assess this knowledge point, and where H is the number.

Let K be a knowledge point that should be evaluated by students in the course. We assume that there are N individual questions $q_k (k = 1, 2, \dots, N)$ associated with knowledge point K . Usually, the questions represent different aspects of K , and then the teacher could know about the degree of acquiring K by aggregating information from those questions. Thus, we define knowledge point $K = \{q_1, q_2, \dots, q_N\}$ as a question set in which the linguistic value of a question expressed with a linguistic 2-tuple (s_k, α_k) defined in Herrera et al. [Herrera and Martinez (2001)]. All options that students can choose for each question q_k can be defined by the notations $S = \{s_0, s_1, \dots, s_M\}$ where $s_i (i = 0, 1, \dots, M)$ stands for i^{th} linguistic term (option) in S and $M + 1$ refers to the

granularity of S .

For example, a set of seven linguistic terms S could be

$$S = \{s_0 = N, s_1 = VL, s_2 = L, s_3 = M, s_4 = H, s_5 = VH, s_6 = P\} \quad (1)$$

where s_0 is a linguistic term which means “Nothing” to some attribute that will be assessed by someone and should be assigned to a numerical value and the others have a different meaning in linguistics. The numerical values of the seven linguistic terms in the set are set to equal to their index values. We can express linguistic information by using the linguistic 2-tuple (s, α) , where s is a linguistic term and α numerical value representing the symbolic translation. Then, we could translate a discrete symbolic term into a numerical value by symbolic translation defined in Herrera et al. [Herrera and Martinez (2000)].

The students are learning the course can play as the experts who will evaluate the teaching quality of some knowledge point by choosing an appropriate term from the linguistic set (4). In the same time, we should distinguish students from each other by crowd grade $\omega_i \in \Omega = \{\omega_1, \omega_2, \dots, \omega_H\}$ which is a comprehensive score computed from multiple grade values expressing some student credibility in the evaluation. In the initiation phase, we should assign all the students with the same crowd grade value 1. At the end of each evaluation process, the crowd grade will be changed according to the student choices.

3.2 Step 2 questionnaire preparation

A teacher should attach some questions behind some slides in digital courseware and then share them with students after they have defined those questions based on real requirements. Generally, those questions are about how students believe they have learned about some known point in a syllabus. At the same time, teachers could decide what kind of students would answer the question. Respondents could be every student sitting in the classroom, someone who scored more than 80 points in mathematics, all the girls, or even ten students selected randomly. For the sake of simplicity, the respondents will be all of the students in this paper.

Optionally, teachers could add some filter questions to mark students who answer “No” to those questions, such as “Have you revised the content of the last lesson?”. Then their questionnaires to specific knowledge points should not be counting or lower their weight.

In a classroom, teachers will play the digital courseware slide by slide in this framework while teaching students. When the last slide to a knowledge point finished, the question will show on the screen of the chosen students. Then all the chosen students will be allowed to answer those questions within the specific time. To limit the answer time could also check whether some students are absent or doing something else in the classroom. After that, all the completed questionnaires will be collected automatically.

3.3 Step 3 symbolic translation

After collecting those questionnaires, we need to express the symbolic information by translating them into a set of linguistic 2-tuple terms.

Definition 1 (Symbolic Translation [Herrera and Martinez (2000)]): Let $\beta \in [0, M]$ be a numerical value aggregating from discrete linguistic information defined in S , where $S = \{s_0, s_1, \dots, s_M\}$ is a linguistic term set, then β can be obtained from the following operation:

$$\Delta : [0, M] \rightarrow S \times [-0.5, 0.5]$$

$$\Delta(\beta) = (s_i, \alpha), \text{ with } \begin{cases} s_i, & i = \text{round}(\beta) \\ \alpha = \beta - i, & \alpha \in [-0.5, 0.5] \end{cases} \quad (2)$$

where the **round** is the normal round operate and α is called *symbolic translation*.

Thus, we can make a linguistic term relationship to a numerical value β and α is the distance to index i as shown in Fig. 2. We also can get the linguistic 2-tuple representing to any linguistic term s_i is $(s_i, 0)$ according to definition 1. At the end of this phase, all the questionnaires will translate into 2-tuple form.

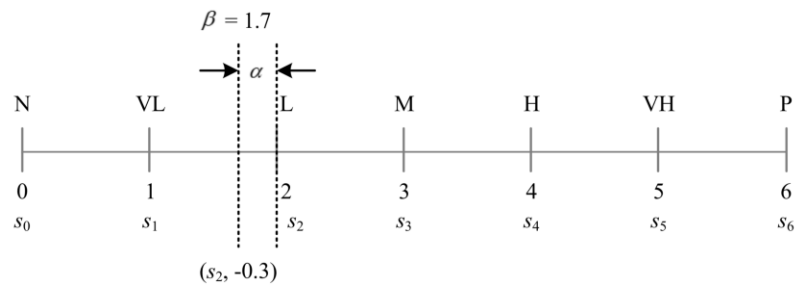


Figure 2: Sample for a representation of a linguistic 2-tuple $(s_2, -0.3)$ value which means it is 0.3 lower than the option “Low” and β is $0.17 = 2 - \alpha$

Besides, the following characteristics are required to be satisfied [Herrera and Martinez (2000)]:

(1) *Comparable*: Let (s_j, α_1) and (s_k, α_2) be two 2-tuples, with each one expressing symbolic information as following:

- a) if $j > k$, then $(s_j, \alpha_1) > (s_k, \alpha_2)$
- b) if $j = k$, then
 - i. if $\alpha_1 = \alpha_2$ then $(s_j, \alpha_1) = (s_k, \alpha_2)$
 - ii. if $\alpha_1 > \alpha_2$ then $(s_j, \alpha_1) > (s_k, \alpha_2)$
 - iii. if $\alpha_1 < \alpha_2$ then $(s_j, \alpha_1) < (s_k, \alpha_2)$

(2) *Negation Operator*: To function Δ over 2-tuple there must be a negation operation $\Delta^{-1} : S \times [-0.5, 0.5] \rightarrow [0, M]$ can be defined per $\Delta^{-1}(s_i, \alpha) = i + \alpha$.

(3) *Aggregation Operator*: Aggregation operator can be used as a mapping from a set of linguistic 2-tuples to one 2-tuple. Many aggregation operators [Merigo, Casanovas and Martinez (2010); Wei and Zhao (2012); Wu, Wu, Zhou et al. (2015); Wei (2011); Mo and

Deng (2016)] have been proposed to adapt to aggregate information in different computation model.

3.4 Step 4 aggregation

Then we will calculate over the linguistic 2-tuples to aggregate the assessment of teaching quality. The different question q_k has a different important factor to describe a knowledge point. To do so, the teacher should assign question q_k a different weight associated ϖ_k indicating how important q_k is to the knowledge point K .

Definition 2 (Arithmetic Mean [Herrera and Martinez (2000)]): Let $q = \{(s_1, \alpha_1), \dots, (s_n, \alpha_n)\}$ be a set of linguistic 2-tuples. Then the 2-tuple arithmetic means can be defined as the following:

$$\bar{q} = (\bar{s}, \bar{\alpha}) = \Delta \left(\frac{1}{n} \sum_{i=1}^n \Delta^{-1}(s_i, \alpha_i) \right) = \Delta \left(\frac{1}{n} \sum_{i=1}^n \beta_i \right) \quad (3)$$

Definition 3 (Weighted Average Operator [Herrera and Martinez (2000)]): Let $q = \{(s_1, \alpha_1), (s_2, \alpha_2), \dots, (s_n, \alpha_n)\}$ be a set of linguistic 2-tuples and $W = \{\varpi_1, \varpi_2, \dots, \varpi_n\}^T$ be their associated weights. Then the 2-tuple weighted average operator \tilde{q} can be defined as follows:

$$\tilde{q} = (\tilde{s}, \tilde{\alpha}) = \Delta \left(\frac{\sum_{i=1}^n \varpi_i \cdot \Delta^{-1}(s_i, \alpha_i)}{\sum_{i=1}^n \varpi_i} \right) = \Delta \left(\frac{\sum_{i=1}^n \varpi_i \cdot \beta_i}{\sum_{i=1}^n \varpi_i} \right) \quad (4)$$

At the beginning of evaluation, all the students have the same *crowd grade* that will define in Step 5 and updated at the rest of loops. Using the crowd grade as for associated weight for each student, we aggregate the weighted average \tilde{q}_k for each question q_k using weighted average operator based on above definition. After all the collective performance values for each question calculated, teachers can get the sorted result about those questions according to the comparable characteristic of linguistic 2-tuple and find out what is the most need to be improved in teaching. Finally, we can calculate the degree of acquiring knowledge point K by aggregating information from all \tilde{q}_k .

3.5 Step 5 update

In this phase, we will update the crowd grade for each student. The crowd grade will be updated at the end of each loop if students will be encouraged to fill in the questionnaire more carefully.

Definition 4 (Distance [Gerogiannis, Rapti and Karageorgos et al. (2015)]): Let (s_i, α_i) and (s_j, α_j) be two linguistic 2-tuples in $q = \{(s_1, \alpha_1), (s_2, \alpha_2), \dots, (s_n, \alpha_n)\}$, then the distance between them can be calculated as the following:

$$d((s_i, \alpha_i), (s_j, \alpha_j)) = |\Delta^{-1}(s_i, \alpha_i) - \Delta^{-1}(s_j, \alpha_j)| \quad (5)$$

Definition 5 (Similarity Degree): Let $q = \{(s_1, \alpha_1), (s_2, \alpha_2), \dots, (s_n, \alpha_n)\}$ and \bar{q} be a linguistic 2-tuples set and the arithmetic mean, respectively, then the similarity degree of any linguistic 2-tuple (s_i, α_i) in q can be calculated as the following:

$$s((s_i, \alpha_i), \bar{q}) = 1 - \frac{d((s_i, \alpha_i), (\bar{s}, \bar{\alpha}))}{\sum_{j=1}^n d((s_j, \alpha_j), (\bar{s}, \bar{\alpha}))} = 1 - \frac{|\Delta^{-1}(s_i, \alpha_i) - \Delta^{-1}(\bar{s}, \bar{\alpha})|}{\sum_{j=1}^n |\Delta^{-1}(s_j, \alpha_j) - \Delta^{-1}(\bar{s}, \bar{\alpha})|} \quad (6)$$

$$= 1 - \frac{\left| \beta_i - \frac{1}{n} \sum_{k=1}^N \beta_k \right|}{\sum_{j=1}^n \left| \beta_j - \frac{1}{n} \sum_{k=1}^N \beta_k \right|}$$

Definition 6 (Accuracy Grade): Let $s((s_{jk}, \alpha_{jk}), \bar{q}_k)$ be the similarity degree of the student j to question q_k , then the **similarity degree** α_j^o of a student j in the loop o for calculating knowledge point K_o can be defined as the following:

$$\alpha_j^o = \frac{\sum_{k=1}^N \bar{\omega}_k \cdot s((s_{jk}, \alpha_{jk}), \bar{q}_k)}{N} \quad (7)$$

where \bar{q}_k is the arithmetic mean of question q_k belonging to knowledge point $K_o = \{q_1, q_2, \dots, q_N\}$.

We assign to each student j an accuracy grade to describe how much better he/she is than average level. The closer to the average mean, the bigger accuracy grade he/she has. The crowd grade of a student is a comprehensive score that can be defined as definition 7.

Definition 7 (Crowd Grade): Let ω_i^o and α_i^o be the crowd grade and accuracy grade, respectively, used to aggregate knowledge point K_o , then the **crowd grade** ω_i^{o+1} used to calculate knowledge point K_{o+1} can be calculated as the following:

$$\omega_i^{o+1} = (1 - \gamma) \omega_i^o + \gamma \frac{L}{N} \alpha_i^o \quad (8)$$

Meanwhile, we merge ω^o and α^o into crowd grade by giving a weight $\gamma \in [0, 1]$ to teaching evaluation for students. The initial value ω^0 of crowd grade for each student set to 1 in our scenario.

3.6 Step 6 improvement

After update phase, the teacher would get the evaluation results generated at Step 4. Thus, the teacher could refine the syllabus and design more effective teaching methods to improve weak to specific knowledge point. At the same time, the achievement of specific knowledge point can also be calculated, and it could be a qualitative indicator of teaching quality. More importantly, the assessment directly comes from students, so it is more reliable and accurate than that from other exports. Moreover, it comes without delay and teacher can get the feedback at the end of each class.

4 Case study

We will illustrate how to use this framework to evaluate teaching quality in this section.

To a knowledge point in the syllabus, the teacher should prepare some questions, such as “Are you interested in this part of the content?”, “Do you understand this part of the content?”, “Do you satisfy this kind of teaching method?”, etc. Then the teacher should set associated weight $\varpi_i \in W$ (with $\sum_{i=1}^M \varpi_i = 1$) to each question that could reflect the different contribution to the achievement degree. In this case, let associated weight set be $W = \{0.2, 0.3, 0.15, 0.2, 0.15\}$. Moreover, the teacher could arrange what kind of students will answer those questions in advance optionally. If not, all students will make feedback by default. After that those questions will be attached to associated slides that can be shared with students in this framework and students will answer them when they are going to be displayed on the students’ screens at the end of that knowledge point teaching.

$$c = \begin{pmatrix} VL & L & L & H & L \\ L & VH & M & H & VL \\ L & H & L & L & L \\ H & H & H & VH & M \\ P & P & VH & P & N \\ M & M & VL & L & L \\ L & VH & L & H & VL \end{pmatrix} \quad (9)$$

We record all collected answers with the matrix c as Eq. (9) shown. According to definition 1, we can translate any linguistic term s_i to the linguistic 2-tuple $(s_i, 0)$. Then we can get matrix C where C_{ij} is a 2-tuple expressing the answer to the question j of the student i (x_i) as shown in Tab. 1.

Then we aggregate all answers to each question q_i from all students by calculating a weighted average to a column i in a matrix C using Eq. (4). Let associated weights be $\Omega = \{\omega_1, \omega_2, \dots, \omega_H\}$ where $\omega_i \in \Omega$ is the crowd grade of a student i with an initial value 1 and that can be calculated by Eq. (8) in the following loops. Thus, we can get aggregated 2-tuples for each question q_i as Eq. (10) shown.

$$(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3, \tilde{q}_4, \tilde{q}_5) = ((M, -0.14), (H, 0.14), (M, -0.29), (H, -0.14), (L, -0.43)) \quad (10)$$

Then the above results can be sorted in increased order and get $(\tilde{q}_5, \tilde{q}_3, \tilde{q}_1, \tilde{q}_4, \tilde{q}_2)$ which means the teacher should pay more attention to question 5 and try to refine it in the following course. Moreover, the achievement for knowledge point K_o can also be calculated by Eq. (4). Let associated weights be $W = \{0.2, 0.3, 0.15, 0.2, 0.15\}$ for five questions and then the 2-tuple representing the achievement of knowledge point K_o is $(M, 0.23)$. The aggregated result means that the teaching quality to knowledge point K_o is in medium level and the teacher may make more effect on this point.

Table 1: Linguistic 2-tuple translated from performance evaluation of all selected students to a knowledge point K_o in $S = \{s_0 = N, s_1 = VL, s_2 = L, s_3 = M, s_4 = H, s_5 = VH, s_6 = P\}$

| Matrix C | | Questions | | | | |
|------------|-------|-----------|---------|---------|---------|---------|
| | | q_1 | q_2 | q_3 | q_4 | q_5 |
| Students | x_1 | (VL, 0) | (L, 0) | (L, 0) | (H, 0) | (L, 0) |
| | x_2 | (L, 0) | (VH, 0) | (M, 0) | (H, 0) | (VL, 0) |
| | x_3 | (L, 0) | (H, 0) | (L, 0) | (L, 0) | (L, 0) |
| | x_4 | (H, 0) | (H, 0) | (H, 0) | (VH, 0) | (M, 0) |
| | x_5 | (P, 0) | (P, 0) | (VH, 0) | (P, 0) | (N, 0) |
| | x_6 | (M, 0) | (M, 0) | (VL, 0) | (L, 0) | (L, 0) |
| | x_7 | (L, 0) | (VH, 0) | (L, 0) | (H, 0) | (VL, 0) |

Finally, let γ be 0.25 as default value, and we update crowd grad for each student according to Eq. (6) and Eq. (7). In this case, we assume that all seven students answered the questions arranged to them. Then the updated crowd grads are:

$$\Omega^{o+1} = (0.31676, 0.33701, 0.33187, 0.32538, 0.28351, 0.32054, 0.33492)$$

The results show that the crowd grade of a student x_5 is lower than others then the student x_5 will decline in influence in the next loop to evaluate another knowledge point.

5 Experiments

Table 2: The answers that twenty-two students (x_1 to x_{22}) given to questions (q_1 to q_{60}) selected from $S = \{s_0 = N, s_1 = VL, s_2 = L, s_3 = M, s_4 = H, s_5 = VH, s_6 = P\}$ associated with 12 knowledge points (K_1 to K_{12})

| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 | x_8 | x_9 | x_{10} | x_{11} | x_{12} | x_{13} | x_{14} | x_{15} | x_{16} | x_{17} | x_{18} | x_{19} | x_{20} | x_{21} | x_{22} |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| q_1 | VL | L | P | H | P | M | L | L | VL | L | L | VL | P | L | M | M | M | VH | L | VL | L | |
| q_2 | L | VH | P | H | P | M | VH | H | M | VH | L | P | VH | M | H | H | H | VH | M | L | H | H |
| q_3 | L | M | P | H | VH | VL | L | L | M | L | L | M | VL | H | M | M | L | L | VL | L | M | L |
| q_4 | H | H | P | VH | P | M | H | H | H | L | VH | M | M | VH | H | M | H | H | H | H | H | |
| q_5 | VL | VL | P | M | N | L | VL | VL | VL | L | L | M | L | L | VL | L | VL | M | L | M | VL | L |
| q_6 | M | H | P | M | N | M | L | H | M | L | L | H | H | M | M | M | L | H | M | M | H | M |
| q_7 | H | VH | P | H | N | VH | VH | H | M | VH | L | P | VH | P | VH | L | H | H | L | M | VH | VH |
| q_8 | H | M | P | M | L | M | M | L | H | M | L | H | M | M | M | H | L | L | H | L | M | M |
| q_9 | VL | L | P | M | M | L | L | VL | L | N | L | N | L | L | VL | VL | L | M | P | VL | L | VL |
| q_{10} | L | VL | P | M | VH | L | VL | VL | M | M | L | L | L | M | M | L | VL | M | P | VL | H | H |

continued...

...continued

| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 | x_8 | x_9 | x_{10} | x_{11} | x_{12} | x_{13} | x_{14} | x_{15} | x_{16} | x_{17} | x_{18} | x_{19} | x_{20} | x_{21} | x_{22} |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| q_{11} | M | L | P | M | VL | M | L | L | H | H | L | M | M | L | M | M | H | L | H | M | L | H |
| q_{12} | L | M | P | L | L | VL | L | M | M | L | L | M | L | VL | L | M | L | M | VH | L | VL | L |
| q_{13} | H | VH | P | VH | M | VH | H | M | VH | M | L | VL | VH | M | H | VH | M | H | VL | H | VH | H |
| q_{14} | P | P | P | P | P | P | VH | VH | H | VH | L | H | H | VH | P | VH | P | VH | VL | H | H | P |
| q_{15} | M | M | P | L | VH | H | L | M | M | L | L | H | L | L | M | M | H | H | VL | L | M | M |
| q_{16} | M | M | P | M | VL | M | H | H | L | M | L | H | H | M | H | M | M | H | N | M | H | H |
| q_{17} | H | M | P | H | VL | VH | H | VH | M | H | L | VH | M | M | VH | M | H | H | N | VH | H | H |
| q_{18} | M | M | P | H | VL | VH | VH | H | M | M | L | VH | P | VH | H | H | M | VH | VL | H | VH | M |
| q_{19} | M | VH | P | VH | L | M | H | H | H | M | L | VH | VH | M | M | VH | P | VH | VH | P | VH | H |
| q_{20} | L | VL | P | L | M | VL | M | L | L | M | L | L | VL | N | VL | VL | L | L | VH | M | VL | L |
| q_{21} | M | H | P | H | L | P | VH | M | M | H | L | H | M | H | H | M | M | L | P | VH | H | M |
| q_{22} | L | VL | P | M | VH | L | L | VL | N | L | L | VL | L | M | L | VL | VL | VL | H | L | L | VL |
| q_{23} | P | VH | P | P | VL | H | VH | VH | H | VH | L | VH | P | VH | P | VH | H | VH | N | H | VH | H |
| q_{24} | L | M | P | M | H | VH | H | VL | VH | H | L | L | M | H | L | H | M | M | VH | M | M | M |
| q_{25} | L | L | P | VL | M | VL | N | N | VL | VL | L | L | VL | L | L | L | VL | VL | H | M | VL | M |
| q_{26} | VH | H | P | H | L | VH | H | H | M | H | L | H | VH | H | VH | VH | VH | P | L | P | H | H |
| q_{27} | M | VH | P | H | H | L | M | M | H | H | L | H | H | VH | M | M | VH | M | L | VH | H | H |
| q_{28} | L | VL | P | L | M | L | VL | H | M | VH | L | M | M | VL | L | L | L | M | L | M | L | M |
| q_{29} | M | M | P | L | VH | M | M | VL | L | L | L | M | L | M | L | L | M | M | VH | VL | VL | M |
| q_{30} | M | M | P | M | L | H | M | L | M | M | L | H | L | M | M | M | H | M | VL | L | M | H |
| q_{31} | L | L | P | VL | VH | L | VL | M | L | N | L | L | L | L | VL | L | VL | L | N | L | VL | L |
| q_{32} | VH | H | P | H | L | H | VH | H | VH | M | L | VH | M | VH | VH | VH | VH | VH | N | H | VH | VH |
| q_{33} | H | H | P | H | L | H | VH | VH | M | VH | L | VH | H | VH | H | VH | VH | M | P | VH | L | VH |
| q_{34} | VH | H | P | H | VL | VH | H | VH | H | VH | L | VH | VH | H | H | VH | VH | M | VH | VH | H | H |
| q_{35} | L | L | P | VL | VH | M | L | M | L | L | L | M | VL | N | L | N | L | N | L | VL | L | VL |
| q_{36} | M | H | P | M | VH | H | L | L | M | M | L | M | M | H | VL | L | H | M | H | M | VH | H |
| q_{37} | M | M | P | H | L | VH | H | H | H | M | L | VH | VH | VH | VH | M | P | VH | L | H | VH | H |
| q_{38} | M | H | P | M | P | M | H | L | H | H | L | M | M | M | H | M | H | L | VH | M | M | M |
| q_{39} | M | M | P | H | P | H | L | M | M | H | L | H | L | L | M | M | M | M | L | H | M | H |
| q_{40} | M | L | P | M | N | M | L | H | H | M | L | M | M | H | M | H | L | M | VH | M | M | M |

continued...

...continued

| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 | x_8 | x_9 | x_{10} | x_{11} | x_{12} | x_{13} | x_{14} | x_{15} | x_{16} | x_{17} | x_{18} | x_{19} | x_{20} | x_{21} | x_{22} | |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|--|
| q_{41} | M | VH | P | H | L | H | H | VH | M | VH | L | H | H | VHVH | M | M | VH | VL | H | VH | H | | |
| q_{42} | P | VH | P | VH | VL | P | VH | H | VH | M | L | M | VH | H | H | VH | M | M | L | VH | P | M | |
| q_{43} | M | M | P | VH | VH | M | H | H | M | L | L | M | M | M | VH | H | M | M | H | M | M | H | |
| q_{44} | L | VL | P | L | VH | L | M | L | VL | L | L | L | VL | VL | N | VL | L | VL | M | N | VL | VL | |
| q_{45} | M | M | P | H | VL | M | M | H | M | M | L | H | H | H | M | H | M | H | VH | H | M | M | |
| q_{46} | H | VH | P | H | P | H | VH | H | VH | M | L | M | H | H | M | VHVH | M | M | H | H | H | | |
| q_{47} | L | L | P | VL | P | L | N | VL | VL | N | L | VL | N | L | VL | N | M | VL | M | VL | L | VL | |
| q_{48} | M | H | P | H | VH | H | M | L | M | M | L | H | H | H | VHVH | VH | H | VL | M | VHVH | | | |
| q_{49} | M | M | P | M | VL | H | M | M | L | H | L | VH | M | M | VH | H | H | H | P | M | H | VH | |
| q_{50} | M | M | P | M | M | H | H | H | H | H | L | H | H | VH | H | H | M | VHVH | VH | H | H | | |
| q_{51} | M | M | P | H | L | M | H | VH | H | H | L | VH | H | H | VH | M | VH | H | P | VH | M | M | |
| q_{52} | M | H | P | H | L | H | M | H | H | N | L | M | H | H | H | M | M | H | VH | M | M | M | |
| q_{53} | M | H | P | L | H | H | M | M | M | M | L | H | M | M | M | H | M | M | L | M | M | H | |
| q_{54} | M | H | P | H | VL | H | H | M | VH | H | L | H | H | H | M | H | H | M | P | H | H | H | |
| q_{55} | L | VL | P | L | H | N | L | VL | VL | VL | L | L | VL | VL | VL | N | VL | N | VH | N | N | VL | |
| q_{56} | H | H | P | H | M | VH | H | VH | H | H | L | VHVH | H | H | VH | M | H | L | VHVH | H | | | |
| q_{57} | M | M | P | M | L | H | H | H | M | M | L | H | M | H | M | M | H | M | H | H | VHVH | | |
| q_{58} | M | M | P | H | H | H | M | M | M | M | L | M | L | M | M | M | M | M | L | H | H | M | |
| q_{59} | VH | H | P | VH | M | P | VH | H | VHVH | L | H | VH | P | VH | P | VH | P | L | VHVH | P | | | |
| q_{60} | VL | L | P | L | L | VL | VL | VL | N | VL | L | L | VL | N | VL | VL | N | VL | M | N | VL | VL | |

To demonstrate the practicability of the proposed approach in this paper, we first developed a prototype with basic functions to evaluate teaching quality. Then we gather a series of questionnaires completed by twenty-two students in one class, and which consist of sixty questions and cover twelve knowledge points of the *Java Programming* course. There are five questions to each knowledge points and totally sixty answers as shown in Tab. 2.

We input these answers into the implemented system manually and get the achievement of each knowledge points. Fig. 3 shows that the evaluation results of the question q_{28} and q_{25} for knowledge point K_6 and K_5 , respectively, are not satisfied, and the teacher should pay more attention to those two aspects in the future teaching.

All the twenty-two students have the same initial *crowd grade* with value one which would be updated at the end of each loop. Fig. 4 shows that there are four abnormal crowd grade curves which represent four students x_3 , x_5 , x_{11} and x_{19} . Comparing them with other students in Tab. 2, we can find that student x_3 and x_{11} simply filled out all the questionnaires with the same answer P and student x_{11} seemed to give the answers

randomly. So their crowd grade is getting lower and lower. The crowd grade of a student x_{19} is the most different with others, and there is a clear rebound from the seventh loop. Analyzing the answers those student x_{19} given, the answers in the first six questionnaires are random that shows the student x_{19} has no intention at the beginning.

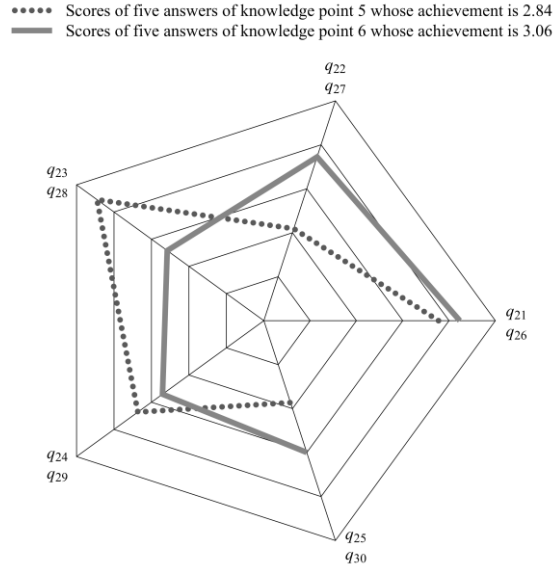


Figure 3: Two evaluation results of knowledge points K_5 and K_6 generated by the implemented system according to 264 questionnaires each of which consists of five questions to the *Java Programming* course and filled out by 22 students of one class in 12 installments

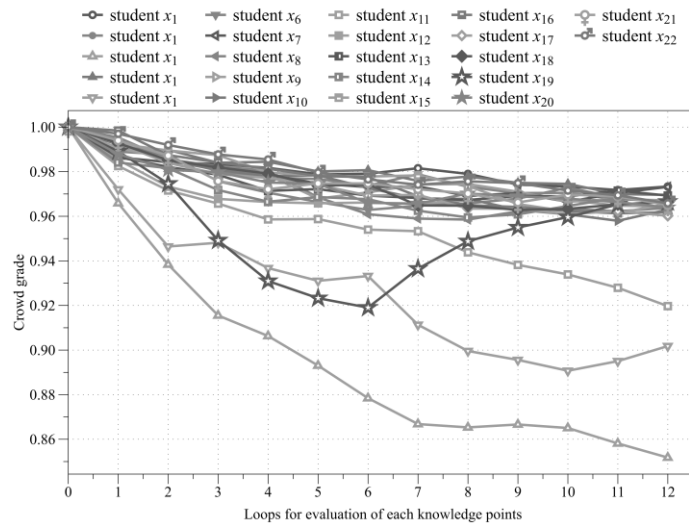


Figure 4: Changes in crowd grades of 22 students with different evaluation loops

The rest answers of student x_{19} in the second half of the course. Teachers could take advantage of in the last six questionnaires are close to weighted average mean, and which illustrated the change analyzing results according to Fig. 4 to identify the abnormal students and do some detail works on them.

As the crowd grade of an abnormal student is getting lower and lower if he/she does not change the behavior, the influence on the assessment of knowledge points will gradually decrease. To illustrate this feature, we do another experiment on two modified data sets. One set of data are modified by picking out four abnormal students that could let us get a real assessment without noising, and the other set of data are modified by removing four normal students randomly in which the noises are still existing. The experiment results show that the deviation from the real value is bigger at the beginning but smaller during the second half of the term (see Fig. 5). The reason is that our approach needs some loops to identify the abnormal students and reduce their crowd grades. The results show that our approach has a strong tolerance for abnormal data.

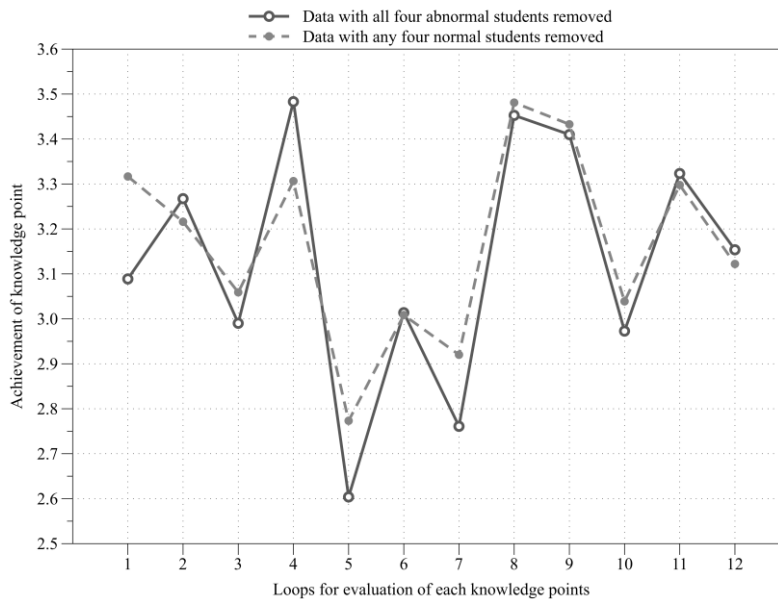


Figure 5: Changes of achievements of 12 knowledge points with two sets of modified data

6 Conclusions

Usually, participants in crowdsourcing have uncertainty during the feedback process, and the uncertainty leads the results obtained from participants may not be accurate. Meanwhile, how to describe and compute this information is a big issue. In this paper, we leverage linguistic 2-tuples to express the feedback from student and proposed a crowdsourcing-based framework to evaluate teaching quality in a classroom. We focus on how to distinguish the difference among all the students. Then we defined crowd grade depending on the accuracy grade to weight ranking from different students. In our approach, student's crowd grade is always updated at the end of each feedback, and that

could encourage students to answer the questions more seriously. Finally, teachers could get the evaluation in time and accuracy could be guaranteed.

Future works will focus on how to describe a student with more properties by using that teacher can send questions to a specific student, and which would make the answers more accurate. If so, we need to develop a mechanism to make sure every student have the similar opportunity to answer the questions.

Acknowledgment: This research is supported by the National Key Research and Development Program of China (No. 2017YFC1502203) and the Key Project of Sichuan Provincial Department of Education (No. 2017GZ0333).

References

Alfaro, L. D.; Shavlovsky, M. (2014): CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, pp. 415-420.

Carrasco, R. A.; Sánchez-Fernández, J.; Muñoz-Leiva, F.; Francisca Blasco, M.; Herrera-Viedma, E. (2017): Evaluation of the hotel's e-services quality under the user's experience. *Soft Computing*, vol. 21, no. 4, pp. 995-1011.

Chang, T. C.; Wang, H. (2016): A multi-criteria group decision-making model for teacher evaluation in higher education based on cloud model and decision tree. *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 12, no. 5, pp. 1243-1262.

Chen, J. F.; Hsieh, H. N.; Do, Q. H. (2015): Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach. *Applied Soft Computing*, vol. 28, pp. 100-108.

Dong, P.; Dai, F. (2009): Evaluation for teaching quality based on fuzzy neural network. *First International Workshop on Education Technology and Computer Science*, vol. 1, pp. 112-115.

Fernandes, J.; Nati, M.; Loumis, N. S.; Nikolettseas, S.; Raptis, T. P. et al. (2015): IoT lab: Towards co-design and IoT solution testing using the crowd. *International Conference on Recent Advances in Internet of Things*, pp. 1-6.

Gerogiannis, V. C.; Rapti, E.; Karageorgos, A.; Fitsilis, P. (2015): On using fuzzy linguistic 2-tuples for the evaluation of human resource suitability in software development tasks. *Advances in Software Engineering*, pp. 1-15.

Herrera, F.; Martinez, L. (2000): A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 6, pp. 746-752.

Herrera, F.; Martinez, L. (2001): A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 31, no. 2, pp. 227-234.

International Engineering Alliance (2014): Washington accord. <http://www.ieagreements.org/accords/washington/>.

Jatoth, C.; Gangadharan, G. R.; Fiore, U.; Buyya, R. (2018): SELCLOUD: A hybrid multi-criteria decision-making model for selection of cloud services. *Soft Computing*, vol. 6, pp. 1-15.

Li, C. C.; Dong, Y. C.; Herrera, F.; Enrique, H. V.; Martinez, L. (2017): Personalized individual semantics in computing with words for supporting linguistic group decision making. An application on consensus reaching. *Information Fusion*, vol. 33, pp. 29-40.

Liu, H. C.; Lin, Q. L.; Wu, J. (2014): Dependent interval 2-tuple linguistic aggregation operators and their application to multiple attribute group decision making. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 22, no. 5, pp. 717-735.

Merigo, J. M.; Casanovas, M.; Martinez, L. (2010): Linguistic aggregation operators for linguistic decision making based on the dempster-shafer theory of evidence. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 18, no. 3, pp. 287-304.

Mo, H.; Deng, Y. (2016): A new aggregating operator for linguistic information based on d numbers. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 24, no. 6, pp. 831-846.

Wang, J. H.; Hao, J. (2006): A new version of 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 3, pp. 435-445.

Wei, G.; Alsaadi, F. E.; Hayat, T.; Alsaedi, A. (2018): Picture 2-tuple linguistic aggregation operators in multiple attribute decision making. *Soft Computing*, vol. 22, no. 3, pp. 989-1002.

Wei, G.; Zhao, X. (2012): Some dependent aggregation operators with 2-tuple linguistic information and their application to multiple attribute group decision making. *Expert Systems with Applications*, vol. 39, no. 5, pp. 5881-5886.

Wei, G. W. (2011): Some harmonic aggregation operators with 2-tuple linguistic assessment information and their application to multiple attribute group decision making. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 19, no. 6, pp. 977-998.

Wu, Q.; Wu, P.; Zhou, Y.; Zhou, L.; Chen, H. et al. (2015): Some 2-tuple linguistic generalized power aggregation operators and their applications to multiple attribute group decision making. *Journal of Intelligent and Fuzzy Systems*, vol. 29, no. 1, pp. 423-436.

Xu, Z.; Liu, Y.; Yen, N.; Mei, L.; Luo, X. et al. (2017): Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, vol. 25, no.1/2, pp. 41-51.

Zhang, X.; Wang, J.; Zhang, H.; Hu, J. (2017): A heterogeneous linguistic MAGDM framework to classroom teaching quality evaluation. *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 13, no. 8, pp. 4929-4956.