Crowdsourcing for Educational Assessment in Large Classes

M.Tech Project Stage I Report
Submitted in partial fulfillment of the requirements for the degree of Master of Technology

by

Raj Agrawal
Roll No: 11305R004

Under the guidance of
Prof. Deepak B. Phatak

Oct 17, 2013

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
# Contents

1 Introduction 1
   1.1 Need of Crowdsourcing in Evaluation 1
   1.2 Peer to Peer Evaluation System 2

2 Crowdsourcing 3
   2.1 Applications 3
      2.1.1 Voting Systems 4
      2.1.2 Information Sharing System 4
      2.1.3 Games 4
      2.1.4 Creative Systems 5
   2.2 Performance 5
      2.2.1 User Participation 5
      2.2.2 Quality Management 5
      2.2.3 Cheating Detection 5

3 Peer to Peer Evaluation System 7
   3.1 Review of Existing Systems 7
      3.1.1 Peer Evaluation Systems for Project Based Learning 7
   3.2 Peer Assessment Procedures for Large Classes 8
   3.3 Challenges 9

4 Proposed System for Peer Evaluation in Large Classes 10
   4.1 Problems in Existing Systems 10
   4.2 Proposed Solution 11
      4.2.1 Measuring Knowledge Level of the Participants 11
      4.2.2 Task Distribution 11
      4.2.3 Allocation Strategies of Projects for Evaluation Based on Knowledge Level 12
      4.2.4 Calculating Final Score 13
   4.3 Summary 14

5 Conclusion and Future Work 15
   5.1 Conclusion 15
Acknowledgements

I would like to express my deepest gratitude to my guide Prof. D.B. Phatak, who has always been making things simple to understand. Without his deep insight into this domain and his valuable time for this project, it would not have been possible for me to move ahead properly. He has been my primary source of motivation and advice during my project. I would also like to thank Mr. Nagesh Karmali for his support and guidance. I also thank Mr. Naveen Bansal and my friends who have always been alongside me throughout this project.
Abstract

MOOC courses like Coursera and recent Stanford’s Online courses have revolutionized learning by online education. Assessment of the participants in these courses is a big challenge, as these courses have huge strengths of participants. Traditional assessment/evaluation approaches are not suitable, as manual evaluation in such a course, having massive participation, is a highly tedious task. In this report, we present an approach to solve this problem by using crowdsourcing. We also present the problems, that arises when peer evaluation is used in large scale, and discussed the problem of fairness in peer evaluation in more detail, and suggested a methodology to address it. Our approach needs to classify participants on the basis of their knowledge levels, later, the tasks are distributed on the basis of these knowledge classes.
Interest in online education is surging and is revolutionized by the success of MOOC courses like Coursera and recent Stanford online courses. For many students, e-learning is the most convenient way to pursue a degree in higher education. A lot of these students are attracted to a flexible, self-paced method of education to attain their degree. Moreover, in asynchronous e-learning classes, students are free to log on and complete work any time they wish. Thus, these students are more likely and more motivated to enroll in an e-learning class.

These MOOC courses have a huge number of participants. These participants (students) have to submit assignments, projects etc. as a part of their course activity for their assessment. The evaluation of these submission in such a large scale is a big challenge in these courses.

1.1 Need of Crowdsourcing in Evaluation

Nowadays, many tasks that are trivial for humans continue to challenge even the most sophisticated computer programs. Prior to the introduction of concept of crowdsourcing these problems were solved by assigning tasks to employees in a company. These tasks can now be done by outsourcing it to a large network of people (crowd). In the field of education instead of minimizing the cost of accomplishing a content creation the objective might be to maximize the expected student competence across a range of topics while holding constant the time spent on videos and worked-problems [1].

Crowdsourcing is a distributed problem-solving and business production model. It is defined as “an idea of outsourcing a task to a large group of people in the form of an open call, that is traditionally performed by an employee” [2].

Since the MOOC courses have a huge number of participants typically in the range of lacs, traditional approaches of assessment and evaluation, in
which the course instructor or TAs or a small groups of people are responsible for evaluation is not feasible. Also automated online evaluation system is not suitable in such a scenario, since these online evaluation system can easily evaluate MCQs but for other patterns of assignments where there are long answers or textual solution these systems are not possible, as one needs to work with lot of NLP and machine learning algorithms. So we can make use of the concept of crowdsourcing for the evaluation of assignments, projects etc. in these courses.

1.2 Peer to Peer Evaluation System

Peer evaluation is the evaluation of work by one or more people of similar competence to the creator of the work (peers). Peer evaluation utilizes the independence and anonymity, of the evaluators in order to discourage cronyism (i.e., favoritism shown to relatives and friends) and obtain an unbiased evaluation. Peer evaluation helps to maintain and enhance the quality of a work by detecting errors and weakness.
Chapter 2

Crowdsourcing

Nowadays, many tasks that are trivial for humans continue to challenge even the most sophisticated computer programs, such as audio transcription and translation. These tasks cannot be computerized. Prior to the introduction of the concept of crowdsourcing, traditional approaches for solving these types of problems focused to assign these task to employees in a company. However it increases the company’s production cost.

Crowdsourcing is a distributed problem-solving and business production model. According to Jeff Howe “crowdsourcing is an idea of outsourcing a task to a large group of people in the form of an open call, that is traditionally performed by an employee”.

Crowdsourcing users can be classified in two categories. They are

- **Requester**: These are the persons who want their work to be done from the crowd. They can also be called as **employer**.

- **Workers**: A Worker is a person who completes work posted by the requester.

The literature on crowdsourcing can be categorized into application and performance.

2.1 Applications

Crowdsourcing applications can be broadly classified in four categories. They are Voting System, Information Sharing System, Games, Creative Systems.
2.1.1 Voting Systems

Amazon Mechanical Turk (MTurk) is a popular crowdsourcing website and voting system. It supports a large number of voting tasks. In MTurk a requester breaks a big task to be done in smaller subtasks. These smaller subtasks are called HITs (Human Intelligent Task). These HITs require a crowdsourcing worker to select his answer from a number of choices. There can be many approaches to get the correct answer from the responses collected from the crowd. One of the popular choice is to consider the correct answer, which is selected by the majority of workers.

2.1.2 Information Sharing System

Websites helps to share information easily among Internet users. Many crowdsourcing systems have been built that aim to share information among the crowd. Some of the examples of such systems are NoiseTube[4], Intelligent Transportation System(ITS)[5]. There are many other famous examples of Information Sharing System such as Wikipedia, Yahoo! Answers etc.

2.1.3 Games

The concept of “Social Game” was introduced by Luis Von Ahn and his colleagues, who created games with a purpose[6]. The games produce useful meta-data as a by-product. By taking advantage of peoples desire to be entertained, problems can be solved efficiently by online game players. The
online ESP Game \[7\] was the first human computation system, and it was subsequently adopted as the Google Image Labeler.

### 2.1.4 Creative Systems

Any advanced technologies cannot replace the role of human in creativity. The creative tasks, such as drawing and coding, can only be done by humans. As a result, some researchers are seeking for crowdsourcing workers to do some creative tasks to reduce the production costs.

### 2.2 Performance

There have been many crowdsourcing applications developed in recent years. Since in crowdsourcing, work is done by very diverse group of people, performance is one of the major factor in crowdsourcing. The performance of the crowdsourcing system can be discussed into following categories. They are

#### 2.2.1 User Participation

In a crowdsourcing system, tasks are distributed to a population of anonymous Internet users for completion. Since completion of task is heavily dependent on the participation from the crowd, it is one of the important factor of the overall performance of the crowdsourcing system. Studies have shown that money is one of the more important motivation of the crowd to participate in a crowdsourcing task. However there are some crowdsourcing systems that don’t offer monetary rewards to their workers. e.g. In Youtube attention measured by the number of downloads, is an important motivation of contributions.

#### 2.2.2 Quality Management

In crowdsourcing the requester of the task break down the task into several small tasks. It is a challenge to decide how the tasks be designed to get good quality output from workers. It was showed that increasing financial incentives, increases the quantity, but not the quality of the crowdsourced task \[8\].

#### 2.2.3 Cheating Detection

Malicious workers often try to maximize their financial gains by producing generic answers rather than actually working on the task. The two approaches to detect cheating workers are: majority decision (MD) and an
approach using a control group (CG) to re-checking the main task [?]. In MD, several workers work on the same task, and the result submitted by the majority of workers is assumed to be correct. In CG only a single worker works on a task, and a control group consisting of other worker re-checks the result for its validity. A task is considered to be correct, if the majority of the control group decides the task is correctly done.
Chapter 3

Peer to Peer Evaluation System

Peer evaluation is the evaluation of work by one or more people of similar competence to the creator of the work (peers). Peer evaluation utilizes the independence and anonymity of the evaluators in order to discourage cronyism (i.e., favoritism shown to relatives and friends) and obtains an unbiased evaluation. It also helps to maintain and enhance the quality of a work by detecting errors and weakness in it.

Peer assessment requires students to use their knowledge and skills to review, clarify and correct others work. These tasks are cognitively demanding and, as they actively engage students with new knowledge, have the potential to reinforce and deepen the understanding of the student assessor. This benefit is particularly apparent if students are involved in evaluating multiple assessment tasks, as they will be repeatedly exposed to the material presented in a variety of formats. Thus, it is often claimed that peer assessment encourages students to become critical, independent learners as they become more familiar with the application of assessment criteria and develop a clearer concept of the topic being reviewed.

3.1 Review of Existing Systems

3.1.1 Peer Evaluation Systems for Project Based Learning

Project based learning at engineering universities are increasingly popular to heighten the attractiveness of engineering degrees and to address professional skills such as teamwork, oral and written communications. In project based learning, students are the part of groups. One of the main challenge in such scenario is that, how to find the contribution of an individual student in the
group. One way is the to use peer and self evaluation in this purpose. Some of the systems built to facilitate such assessments have been discussed

3.1.1.1 Delft:PeEv

PeEv was developed by Delft University of Technology. In PeEv each student in a project group evaluates all other members of their group including themselves. Evaluation is done based on some fixed set of criteria. Student marks their group members on each criteria on a certain scale. PeEv then calculates the average of marks give by group to a student, as his score. The feedback is given to student is graphical or using the text phraseology in terms of Excellent, Good, Satisfactory, Marginal and Poor. Students never see the actual numerical score. The final score is only computed after all students have done the evaluation.

The project group tutors can also evaluate student on the basis of same criteria, on which his peers evaluated. After all students in a project group have scored each other the system displays the results to the students. They can then see per learning objective how others view them compared to their assessment of their own capabilities.[9]

3.1.1.2 SPARK

SPARK is an acronym for Self and Peer Assessment Resource Kit aimed has been developed by the University of Technology, Sydney. The motivation behind the development of spark was same as that in PeEv, that is students get equal group marks for unequal contribution. Free riders and plagiarizers receive a mark that does not represent their efforts and at the same time better students are usually marked down. In Spark there is an option to negotiate the criteria with the students. Spark also follows similar evaluation strategy as PeEv i.e. evaluation based on certain criteria. However in Spark evaluation scale can be adjusted. In Spark, instructor can decide to publish final result or not.[9]

3.2 Peer Assessment Procedures for Large Classes

Most of the literature on peer assessment focuses either on the evaluation of individual contributions to group assignments or on the validity and reliability of such forms of assessment. Few studies have been concerned with evaluating the experience from the viewpoint of students themselves [10]. Furthermore, almost all studies to date involve single tutorials or small classes taught by staff who are committed to the procedure. As in MOOC
courses, classes are of very large size, the development of viable procedure for the use of peer assessment in large classes is important.

A study was conducted in Queenslund University of Technology, to develop the procedures for peer assessment in large classes (not as large as in MOOC). The study was conducted in three phases over the period of two years. Both the student and staff had participated in the project. One assessment task was selected by the coordinator to be included in the study. The assessment itself contributed a proportion of student’s total, ranging in 15-30%.

3.3 Challenges

- Diversity in intelligence and domain knowledge, among participants, creates an unavoidable problem of ensuring fairness in evaluation. Fairness here refers to the ability of participants to evaluates their peers’ assignments. This problem becomes severe, as the number of participants increases.

- Many students feel that students (their peers) are either easy or hard markers. They feel that their score doesn’t reflect the efforts that they have put in the course [10].

- Some feels that peer assessment is a time consuming process and evaluation should be the responsibility of the instructor. However it is shown that a ‘reasonable’ number of marks (typically 10-15% of the total) can be allocated to student performance in the peer evaluation process, as it may boost engagement and commitment to the task [10].

Since there have not been much work done for the peer evaluation system in massive class strengths. We propose a design of the system in the next chapter, that can be developed for this purpose.
Chapter 4

Proposed System for Peer Evaluation in Large Classes

In chapter 3, various existing systems for crowd based peer evaluation are introduced. It is noted that, most of such systems focused on the evaluation of individual contributions in group participation and does not posses the capabilities to handle peer evaluation in large classes. We have also studied the challenges that occur to develop procedure for peer assessment in large classes. Section 3.3, presents a brief introduction of the major challenges encountered while developing a large crowd based evaluation system. There is a need to address these issues in order to construct an evaluation system which is fair to all the participants. It is identified that, distribution of assignments among peers poses major challenge in evaluation. An ideal peer evaluation system should be fair, in order to avoid any discrepancies and wrong perceptions in the participants minds. A fair peer evaluation system will encourage participants to participate in evaluation and increases their beliefs in the system.

In the next section, we address this issue more elaborately and propose a possible solution to cater this issue.

4.1 Problems in Existing Systems

Diversity in intelligence and domain knowledge, among participants, creates an unavoidable problem of ensuring fairness in evaluation. Fairness here refers to the abilities of participants to evaluates their peers’ assignments. This problem becomes severe, as the number of participants increases.

This is one of the major problems in peer evaluation system. The problem of ensuring fairness in peer evaluation system may be resolved by exploiting the knowledge levels of the participants.
4.2 Proposed Solution

The problem of ensuring fairness in peer evaluation system may be resolved by exploiting the knowledge levels of the participants.

4.2.1 Measuring Knowledge Level of the Participants

The knowledge levels of the participants can be measured by several ways e.g.

- When participants interact with the system, they attempt quizzes and exercises during learning of the course. This interaction history can be used to estimate the knowledge level of the participants. This approach assumes that all the data related to performance of the participants is prerecorded and is available for all the participants in the course. However, this may not be possible in all the scenarios. e.g. in case when quiz has not been attempted by all the participants.

- Knowledge levels of the students can also be measured by giving them a quiz on the basis of the task assigned to them for peer evaluation. This approach is intuitive but, in this scenario, the system always depends on an external entity who design the quiz, and the fairness is highly dependent on the quality of the quiz.

- As the above approaches have their own limitations, we can use the combination of these approaches to measure the knowledge level of the participant. This strategy assumes an expert is available. The expert has an option to choose either of the above strategies to measure knowledge level of the participants. If he finds that sufficient performance information of participants in course is available, to measure their knowledge level then he uses this information to measure the knowledge level of the participants. Otherwise, knowledge level is measured by giving them a quiz on the basis of task assigned to them for evaluation.

4.2.2 Task Distribution

In MOOC many of the quizzes and assignments are online and have MCQ type question. These pattern of exams does not requires peer evaluation system, as automated evaluation of these exams can be done fairly. However, there are instances when there are assignments/projects having long answer type solutions. Evaluation of these assignments/projects is not feasible by any automatic systems. For such pattern of assignments/projects human
intervention is required and in such scenarios our peer evaluation is going to be used.

Distribution of tasks (projects/assignments) for evaluation in such a way that can ensure fairness in the evaluation, is the major problem in peer evaluation. We have also studied some approaches to address this problem. In our work, we are considering task distribution as the significantly large problem that needs to be solved. We are proposing different approaches that needs to be tested for the suitability in the real environment.

Continuing with the idea of measuring knowledge levels, there can be many approaches that can be followed to measure knowledge levels of the participants in an efficient ways. An approach for measuring the knowledge level was also discussed in the previous section. We tried to overcome the limitations associated with this method. When the number of participants is large, the diversity among students on knowledge levels is also large. To ensure the fairness in such a diversified crowd of participants, we need to follow a specific approach. One approach in this direction says that, the participants can be divided in the classes, on the basis of knowledge levels. Let we have \( n \) classes, where each class \( C_i \) represents a particular knowledge level. If we divide the students in this set of classes, then the assignment task for evaluation becomes easier. There are many approaches with some limitations associated with each one. One important thing that can be noted here is that, the classes are in the order of knowledge level e.g.

\[
\text{KnowledgeLevel}(C_1) > \text{knowledgelevel}(C_2)
\]

Hence, the participants who belong to class \( C_1 \) posses higher knowledge then the participants of class \( C_2 \).

### 4.2.3 Allocation Strategies of Projects for Evaluation Based on Knowledge Level

Considering the knowledge levels classes of all the participants, the projects can be distributed by one of the following approaches:

- Project of any participant \( P_i \) of class \( C_k \) can only be allocated to the participant \( P_j \) of the class \( C_k \). The motivation behind this approach is on the fact that, participants belonging to same class have same knowledge level, hence, may be able to understand the assignments of their peers belonging to same class in a better way.

**Limitation:** The problem that may arise with this strategy is, participants tend to give higher marks as the evaluator is having the same knowledge level, as they may not be able to find out the flaw in the assignment. Also with this strategy the students might not gain any significant learning advantage as knowledge of levels of both the peers are same.
Allocate the project of participant belongs to class $C_i$ to the participant of $C_{i+1}$. The motivation behind this approach is that, with this approach, participants may have the learning advantage, as the participants will evaluate the assignments of their peers having higher knowledge level compared to themselves.

**Limitation:** This approach may fails, as this does not solve out basic problem of fairness in evaluation, since evaluators might not have sufficient knowledge to do the evaluation.

- Allocate the project of participant belongs to class $C_i$ to the participant of $C_{i-1}$. This approach may ensure the fairness in the evaluation, as the evaluators may be having enough knowledge to evaluate assignments of the peers, whose assignments they evaluate.

**Limitation:** This approach does not solve the problem of learning. As the students evaluating the assignments have higher knowledge level as compared to their peers whose assignments they will evaluate, they may not have any learning advantage, from the process of evaluation.

It should be noted that, each strategy possess their own benefits as well as drawbacks, which does not enable us to use any of them. However a combination of these strategies can solve the problem.

Allocate projects for evaluation of participant $P_i$ of class $C_k$, to exactly 1 participant from each class $C_1, C_2, \ldots, C_n$.

### 4.2.4 Calculating Final Score

**Normalization of scores from each class**

The score given by a participant of class $C_i$ is normalized by multiplying the score by the normalizing factor of that class.

Normalizing factor of class $C_i$ is calculated by taking the ratio of average marks allotted by participants of all the classes to the average marks allotted by the participants of class $C_i$.

$$normalizing\_factor(Class(C_j)) = \frac{\left(\sum_{i=0}^{n} S_i\right)}{S_j}$$

where,

- $S_i$ is the average marks given by the participants of class $C_i$,
- $n$ is the number of classes representing different knowledge level

$$normalized\_score(participant_i) = score(participant_i) \times normalizing\_factor(Class(C_i))$$
where \( score(participant_i) \) = score given by participant of class \( C_i \)

The final score of a particular project is calculated by taking the normalized score from every class for a particular project. The average of this normalized score from all the participants, for a given project, is be considered as the final score for the project.

\[
Score(project(p_j)) = \frac{\sum_{i=0}^{n} normalized\_score(participant_{ij})}{n}
\]

where,
\( n \) = number of participants, participating in the evaluation,
\( normalized\_score(participant_{ij}) \) = normalized score given by participant of class \( C_i \) to project \( p_j \)

### 4.3 Summary

In this chapter we presented the problem of fairness in peer evaluation. We have proposed that, we can classify the participants in different classes based on their knowledge level. On the basis of this classification we can distribute the task of evaluation to the participants to ensure fairness. The final score of the evaluation is calculated by taking the average of normalized score given by participants.
Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this report, we have seen why we need the use of Crowdsourcing for education assessment in large scale. We have also studied various types of crowdsourcing system and performance issues related to them. We have reviewed existing peer to peer evaluation systems and problems related to them and found that fairness in evaluation is major problem. By exploiting the knowledge level information about the peers this problem can resolved to a certain extent.
Bibliography


