ABSTRACT
We provide experimental results that help automate two related tasks around online instructional forum use. We show how forum posts can automatically be evaluated for inclusion in a course specific frequently asked questions list. We further show how computer algorithms can help instructors identify students who deserve course credit for their participation in the class forum. We deploy forum usage statistics that are readily available in most online forum facilities to train two algorithms. For the FAQ inclusion task we train a logistic regression to an accuracy of 0.78. As ground truth we asked teaching assistants of one course to provide expert judgment for use in the regression training phase. For the academic credit apportioning task we again used experts to compare pairs of forum contributions. Using those judgments we trained a support vector machine to approximate the resulting rankings of posts. We reached a ranking accuracy of 76%, where accuracy reflects minimization of faulty pairwise rankings compared to the expert gold set.

ACM Classification Keywords

Author Keywords
Forum; Forum credit; MOOCs; residential courses; FAQ; online learning; social network analysis

INTRODUCTION
As institutions of higher education seek to expand their geographic coverage, online forum facilities have been the most obvious tool at hand. Massively open online classes (MOOCs) deployed them immediately along with the medium of instructional video.

Ideally, an online forum allows students to help each other, and to communicate with instructors and course assistants. In the case of geographically distributed learning populations, no time of physical collocation for discussion is available.

The forum has then been a main means for information flow outside the unidirectional video stream.

Yet it is not only in the context of distance learning that forum facilities have found uses. Even when in person class time is available, many residential college courses have adopted the tool. The need for students to ask questions, voice concerns, or to point out errors in course material are as salient in residential settings as they are in less traditional situations, such as distance learning.

Figure 1 shows the increase in contributions to just one of several available online forum tools over the past years in a large private university. Several hundred residential courses make heavy use of the facility. The Related Work section below points to some of the reasons for this popularity.

In fact, the perceived benefit of forum participation for students has been high enough that instructors are awarding between 1% and 25% credit for students who are active online throughout the term. This trend may well continue as courses for which extended discussions are themselves core to the learning begin to adopt the tool.

Several years worth of a course’s forum archive hold treasures for a number of stakeholders. The answers to many questions might be buried in those archives, and would save time for future students and teaching staff alike—if those posts were available.

Instructors taking over the teaching of a course from a colleague could theoretically learn from those archives where students tend to falter. In fact, we can think of forum con-
We present experiments that evaluate methods for automatically generating frequently asked question lists (FAQs). The latter have the advantage that they can be honed over time to maximize utility. Students themselves could in fact participate in the improvement of the lists.

Who, however, can take the time to comb forum archives for answers worth including in a specially curated list? An automated construction and maintenance of frequently asked question lists from live and archived forum posts is a more promising proposition. Yet the question is how an automated process can identify ‘important’ forum items.

Natural language processing comes to mind (NLP). But courses are of vastly differing domains. That variety is not easy to conquer using NLP. And what is highly important in the moment, like a typo in a high credit problem set might be of little interest down the road. Automated FAQ content selection through semantic understanding is therefore a difficult problem.

In addition, the ‘quality’ of a question or answer can be subjective. What may appear to an instructor as an obvious student answer to a peer’s online question may in fact be a strong contribution to the class. Unbeknownst to the instructor, others may be just as confused as the student posing the question. The instructor would thus not think of curating the question/answer pair into a FAQ list.

Inversely, a note from an instructor that is not recognized by students as important may in fact be of high value. Neither instructors nor students alone are therefore necessarily solely competent curators of a FAQ list.

The lure of academic credit to incentivize online student participation brings its own difficulties. One problem is that some students unfortunately try to game the system by producing forum entries from marginally modified existing contributions by peers. A more interesting issue is related to the FAQ selection problem: Not every forum contribution is worthy of credit. Volume alone should not be a satisfying criterion.

At the end of an academic term instructors and their assistants thus face the daunting task of evaluating which student actions deserves more credit than others. Instructors could scan the term’s worth of posts, or randomly sample from the collection. For a large class even these partial solutions are expensive. Some instructors cobble together ad hoc measures from statistics that are available as part of forum facilities. These statistics are in fact useful. But they need to be utilized with guidance from empirical evidence as to their relative value.

We present experiments that evaluate methods for (i) automatically generating frequently asked question lists, and (ii) help instructors identify worthy recipients of forum participation credit.

We embed these facilities in a forum centric software tool that is intended to be expanded over time. Figure 2 shows a block diagram of the current system. We focus here on the FAQ generator and grading aid, but briefly mention our current spam detection facility at the end. As represented in the Figure, we ensure that instructors can customize the credit proposing engine to their preference.

We will show that machine learning algorithms in conjunction with available forum statistics can reveal which student actions on a forum are strong indicators of either post importance, or credit worthy contribution. Given those results we can deploy the same algorithms to accomplish these tasks on an ongoing basis.

After a brief literature review we introduce the statistics that are available in the forum facility on which we focused. We then turn to details of our experiment around FAQ list candidate detection. After presenting results, and a discussion of an associated experiment, we move on to the credit apportioning task. Our solution to this problem also rests on the results of an experiment, whose results we describe and discuss. Short coverage of the spam detection, and a conclusion close out the paper.

**RELATED WORK**

Over the years, online discussion forums have become a primary focus of educational research [12]. Forums provide a fertile ground for collaborative learning and engagement in online courses [1]. Many researchers have strongly endorsed the importance of interaction in collaborative learning. This interaction captured in the transcripts of the threaded discussions is an object of active research. In [3], the authors present an overview of different content analysis instruments to analyze transcripts of asynchronous, computer mediated forums in formal educational settings. In the initial stages, the focus was to gather quantitative data to gauge the level of student participation. However, gradually, content analysis began to be adopted as a technique to unravel the information captured in these interactions.
In [16], J. Wong et. al present an analysis of MOOC discussion forum interactions from the most active users. The authors identify the most active participants or posters in the forum discussions and study their effect in the learning environment of discussion forums. Using three simple metrics: views, replies and duration of a thread, they arrive at the conclusion that the most active users are also the most influential users in online forums. Furthermore, they used positive and negative votes on posts and comments to arrive at the result that the active users generally make a positive contribution to the forum.

In [13], the authors demonstrate the use of social network analysis in analyzing information networks embedded in e-learning environments, in order to help the instructors assess the participation of students. [13] relies on assessing the participation of student separately using two types of networks - interaction network between students in a course and the network of terms used in their interactions. The dynamic visualization of interaction between participants and the groups or communities formed can help the instructors rank students based on their centrality in the students’ interaction network. Visualizing the network of terms used in an online discussion forum can be used to compare the interest of different students and their relative engagement.

In contrast with the aforementioned works which focus only on certain specific assessment criteria (either quantitative or using social network analysis), the authors’ approach for assessing a student’s contribution uses a weighted combination of qualitative measures, quantitative measures, engagement level measures and measures from social network analysis, in order to provide a holistic view of the student’s contribution. In order to create the ground truth, the authors rely on expert human judgment. Then, the model is trained to tune the individual weightage of each of the features used in the system. This is the default configuration provided to the instructors. The system also allows the weightage of each feature to be configurable by the instructors, thereby giving them complete autonomy over the grading criteria for their course. Not only this, the system also has a spam detector component in order to flag the students who try to trick the system, so that the instructors can factor that into the grading as well.

Since the discussion forums possess a wealth of information that can be put to great use for subsequent offerings of the course, in this paper the authors propose a system for automatic FAQ creation using the forum postings of a course (collected over the years). Most of the research in the FAQ world is focused on FAQ retrieval rather than automatic FAQ creation. The work by Hu et al. in [5] proposes a semi-automatic method to identify similar questions posted in Open Source Projects forum in order to assist the managers in constructing the FAQ. In [17], the authors proposed a knowledge share platform structured as a FAQ, which is constantly enriched by the newly generated interactions in the community. More recent work by [10] is similar to [17], with the difference that the FAQ managers in [10] also have the usage information at their disposal to monitor the FAQ performance and tailor it based on their needs. The related works discussed above are limited in one or more of the following ways:

- Need for extra effort for creation and maintenance of contextual keywords vocabulary.
- FAQ creation is semi-automatic and needs the intervention of FAQ managers.
- Relies on the feedback of the users for the retrieved FAQ to maintain the FAQ quality.

The work by the authors of this paper addresses all the above limitations and is vastly different from the previous works in the area. In contrast with the above approaches, the authors’ approach is heavily focused on building FAQs using forum posts from the current and previous offerings of the course. Statistical methods (using features like number of upvotes on the question, number of views on the question, number of unique collaborators in the thread) are used to filter out some potentially useful question-answer pairs from the entire dataset. Expert judgement of humans is then used as ground truth to train the system to be able to automatically filter out the most useful FAQ question-answer pairs from the huge dataset. There is no need for any contextual keywords vocabulary for this system. Besides, the FAQ creation is a completely automatic process, requiring minimal or no manual intervention.

BACKGROUND

Many of our classes use the Piazza forum facility [14]. In order to understand the experiments below, we provide a brief overview of this tool.

Piazza is a Q&A web service for online discussions, where users can ask questions, answer questions, and post notes. Every post can be organized into pre-defined folders (e.g. assignment1, logistics, etc) for easier search. The user interface contains a dynamic list of posts, which are question titles followed by a snippet of lines from the post. The posts are arranged chronologically on the left panel, while the central panel is meant for viewing posts and adding contributions. For every question there is a placeholder for the instructor’s answer, which can only be edited by the instructors. There is also a students’ answer section where students collaborate to construct a single answer. A version history is maintained for the students’ answer to identify the specific contributions made by each student. Students can upvote each others’ questions or answers. Instructors can also endorse good questions and answers, which are then highlighted as instructor endorsed. There is also a discussion segment for follow-up threads.

Piazza provides the instructors with certain statistics for each student. These statistics are: number of questions asked, number of questions answered, total number of contributions (including posts, responses, edits, follow-ups etc.), the number of days the student was online, and the total number of posts viewed by the student.

Apart from the readily available statistics from Piazza, the authors used a few information sources, which are described in the upcoming sections. Machine learning techniques were used to select the most powerful among these features, and to tune their weightage. The next sections describe the authors’ subsystems and experiments in detail.
AUTOMATIC FAQ GENERATION
The first subsystem addresses the problem of automatically identifying forum question–answer pairs that are appropriate for inclusion in a class FAQ list.

We worked on the creation of two separate FAQs, each for a different use case:

- Student FAQ: This FAQ is intended to include mostly conceptual questions. Given that the core concepts covered in the class change very little year on year, the concepts that students struggle with have significant overlap over the years. Hence, this FAQ can increase their productivity by providing them instant help without having to wait indefinitely for an instructor’s response. For the course staff the FAQ translates to reduced workload.

- Instructor FAQ: This FAQ is meant to serve as a snapshot of the previous offerings of the course. The information is to help new or continuing instructors take corrective actions to improve the student experience over the years. The FAQ thus focuses on the topics that the students struggled with in the past, and on logistical mishaps of earlier offerings.

We can only cover the first of these two facilities.

To obtain ground truth for training machine learning algorithms, human experts were consulted. A survey format was used, in which the experts were presented with a series of forum contributions (question–answer pairs and notes). The next section discusses how the authors chose the contributions to be included in the survey, and created the ground truth used for training.

Candidate FAQ set and collection of ground truth
We began with the complete dataset of contributions from four years of a graduate level course that introduces artificial intelligence (AI). From this set, contributions were culled to ensure that all the candidate training features we planned to use were represented in the set. The features of interest are as follows:

- Number of upvotes on questions or notes
- Number of upvotes on students’ answers to questions
- Number of upvotes on any instructors’ answers
- Number of unique collaborators in a thread
- Number of unique views
- Length of follow-up threads, or number of follow ups.

Individual threshold percentiles were set for each of the above features. These thresholds served to assemble the candidate set for the expert opinion surveys. For instance, a 60th percentile threshold on the number of question upvotes would mean that to be included in the survey a question would need enough upvotes to clear that hurdle.

The thresholds for each of the above features were manually tuned such that a few hand-picked important questions survived. These thresholds were increased as we progressively selected from older datasets. It was assumed that more recent offerings would have posts that are more relevant to upcoming offerings.

Beyond the post samples selected as above, some additional random samples were included in the survey.

These samples were then presented to the experts in random order. In total, 75 question–answer pairs and notes were selected for the survey.

In order to avoid fatiguing our experts, the set of 75 items was partitioned into three batches. These surveys were administered such that each question–answer pair or note was seen by at least three experts. One of the batches had the benefit of a larger panel of 5 experts. The eleven experts were recruited from among the current course staff for the same class from which the forum contributions were drawn.

The survey instructions were as follows:

Please choose ‘Yes’ or ‘No’ for each of the following questions/notes.
A ‘Yes’ indicates that the respective forum item would be relevant and useful in a FAQ list for the class.
A ‘No’ means that the item should not be included in the FAQ for future offerings of the class.
The entries in the FAQ are taken from previous iterations of the class. Their intent is to answer student questions instantly, and reduce TA workload (by answering common student questions)

Two sample items from the survey are as follows:

**Question:** For the definition of Markov Blanket does it refer to finding the neighbors of A in terms of the factor graph or in terms of the Bayesian network

**Response:** The Markov blanket refers to all variables that have a factor in common with at least one of the variables in A. If two variables are connected in a Bayesian Network, then they must share a factor in the corresponding factor graph representation.

**Question:** I just want to make sure I understand what exactly the hidden cases mean. They are cases that we don’t know the correct answer to, but we can be sure that if it there is no time out/crash on our hidden cases, then it should be fine (in terms of not failing for time outs/crashes) when it gets graded?

**Response:** Yep! You can even doubly verify this by running the grader.py script on corn, so that you can verify that even on [our] machines it doesn’t crash or time out.

Once the survey yes/no responses were available, we deployed them for training. The next subsection introduces details of this process.

**Experiment: FAQs for Students**
Given the binary expert decisions we constructed three types of binary include/exclude scores for each forum contribution. The purpose of this variety was to then identify a machine learning model that would best mimic our expert panel. We thus derived three one-dimensional response vectors from the surveys for use in training.
We trained three logistic regression models, presenting the results of their performance in Table 1. We examined how much the experts agreed with each other concerning whether a forum contribution should be included in the FAQ. As explained, the predictors whose weights the training needed to optimize were:

- Number of item views
- Number of unique collaborations in the thread
- Number of upvotes on student answers
- Number of upvotes on instructor provided answers
- Number of upvotes on the question
- Number of item views

We computed Fleiss' kappa at 0.20 at \( p < 0.5 \). These disagreements stem from the natural subjectivity involved in constructing a FAQ.

The numbers show more detailed results of the three logistic regression models. Numbers for the best performing model are bold faced.

<table>
<thead>
<tr>
<th></th>
<th>Majority</th>
<th>UnanimousYes</th>
<th>Unanimous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.78</td>
<td>0.8</td>
<td>0.64</td>
</tr>
<tr>
<td>95% CI</td>
<td>(0.64, 0.88)</td>
<td>(0.66, 0.90)</td>
<td>(0.49, 0.77)</td>
</tr>
<tr>
<td>Precision</td>
<td>0.74</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.97</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>F1</td>
<td>0.84</td>
<td></td>
<td>0.18</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.5</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>( p ) [Acc&gt;NIR]</td>
<td>&lt; 0.1</td>
<td>&gt; 0.5</td>
<td>&gt; 0.5</td>
</tr>
<tr>
<td>McNemar's ( p )</td>
<td>&lt; 0.5</td>
<td>&lt; 0.1</td>
<td>&lt; 0.1</td>
</tr>
</tbody>
</table>

Table 1: Performance of three logistic regression models

We did not anticipate its strong showing, hypothesizing that inclusion of such answers was often of short term interest to a particular course offering.

The values of all coefficients are zero on the left, being completely suppressed by the large penalty term. As the L1 norm increases, more of the predictors come into play. The dotted vertical line shows the optimal lambda for control of the penalty term to be 1.0.

Table 2 shows the coefficients chosen for the predictors in the majority model. Note the predictors ked not used. They correspond to the lines that still cover the horizontal zero line at \( x = \lambda \) in Figure 4. The table includes the model’s confusion matrix.

Discussion: FAQs for Students

The majority model’s confusion matrix shows good performance identifying forum contributions to be included. On the other hand, admission to the FAQ is overly promiscuous, as witnessed by the ten items that are included, but were excluded by the expert panel. It is possible that this deficiency stems from an imbalance in the training set. The experts included 1.5 times as many posts from their sample as they excluded. The training phase therefore did not encounter as many negative examples as it did for inclusions.

It is surprising that upvoted answers by instructors did not have a strong enough predictive impact to be included in the model. It is possible that such answers were often of short term interest to a particular course offering.

The strong participation of item view frequency is interesting. We did not anticipate its strong showing, hypothesizing that active steps, such as upvoting would be strong contributors to FAQ eligibility. In retrospect, views are the lowest friction method for students to ‘vote with their eyeballs.’
Table 2: Coefficients and confusion matrix for the best FAQ inclusion decision model

<table>
<thead>
<tr>
<th>Predictor coefficient</th>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert Panel</td>
</tr>
<tr>
<td></td>
<td>Prediction</td>
</tr>
<tr>
<td>Intercept -0.42770771</td>
<td>Include 29</td>
</tr>
<tr>
<td>Upvotes on questions -0.16599701</td>
<td>Exclude 1</td>
</tr>
<tr>
<td>Upvotes on instructor answers not used</td>
<td>Include 0</td>
</tr>
<tr>
<td>Upvotes on student answers -0.09142915</td>
<td>Exclude 4</td>
</tr>
<tr>
<td>Number of collaborations not used</td>
<td>Include 0</td>
</tr>
<tr>
<td>Number of item views 0.54748083</td>
<td>Exclude 0</td>
</tr>
</tbody>
</table>

Figure 4: Changes in coefficients as the L1 norm changes.

ASSESSING FORUM PARTICIPATION

Forum participation is increasingly becoming an integral component of the grading policy for a course. The study in [18] indicates that the students’ forum participation increases as instructors include forum contribution in their course grading component. This section focuses on the system built to apportion fair credit to the students for their forum participation.

Limitations of current grading schemes

On surveying several instructors who use forum participation in their grading scheme, the authors found that most of the instructors rely on the basic quantitative statistics that the forum machinery readily offers. The following were some of the grading schemes that are currently used by instructors at the University for awarding forum participation grades:

Scheme 1: In this scheme, scores of each student were calculated using the following formula:

\[
Score = 1 \times \text{numberofquestionsasked} + 4 \times \text{numberofquestionsanswered} + 0.5 \times \text{othercontributions}
\]  

Scheme 2: In this scheme, scores of each student were calculated using the following formula:

\[
Score = 3 \times \text{numberofquestionsanswered} + 1 \times \text{numberoffollowups}
\]  

Anyone above the 90th percentile received full credit, and all the other students received a score of 0.

Scheme 3: Award full credit if at least one contribution was made, and the student was online on the forum for at least x number of days, or viewed at least y posts. Here x and y were set by the instructor using intuition.

Scheme 4: Award full credit to a student if he/she made at least one contribution to the forum.

All the above methods rely solely on the basic statistics directly provided by Piazza. The concern, however is whether these methods accurately and meaningfully award credit to the deserving student. Following are two major limitations of using the current grading schemes:

- Lack of qualitative measures: All the schemes above overlook the quality of contributions. This exclusion negatively impacts the grades of the students who post few, but very high quality contributions. More importantly, relying solely on the quantity of the contributions encourages posts that do not constitute meaningful forum participation. This behavior in turn can cause the forum’s quality to devolve.

- Reward not proportionate to effort: Most of these schemes do not award credit proportionate to the amount of effort and time the student invested. For instance, using the third and fourth scheme meant that two students with vastly differing quantity and quality of contributions would be awarded the same score. Concretely, let us consider two students A and B. Student A made only one forum contribution during the course by posting a "+1" to another student’s question. However, student B regularly made meaningful forum contributions throughout the quarter. Using grading scheme 4, both would receive equal credit. This lack of fairness can deter students from engaging in meaningful forum contributions.

Despite the above limitations, the instructors have no choice but to rely on grading schemes like the ones discussed above. The large volumes of forum posts that accumulate by the end of the term make it impossible for the course staff to manually go through them to apportion credit. However, even if hypothetically, one were to have the course staff manually go through each of the contributions, there is a significant amount
of subjectivity in assessing forum contributions. Having TAs manually grade contributions would lead to a lack of grading uniformity. A trivial contribution according to one TA, might be a significant contribution to another. Thus, there is a need for an automated way to assess the forum participation of students using a holistic grading scheme. Automation can lead to a standardized approach across the entire class.

The next sections discuss how the authors developed a system to assess forum participation by each student in order to award them fair credit. The work moves beyond the ready at hand usage statistics. The methods add insight into the dynamics of student involvement in the forum. These dynamics manifest in the social networks that are created by the online interactions. We briefly introduce the concepts, because they contribute to our set of predictors.

**Extracting Social networks from forum interactions**

Social network analysis examines the structure and composition of ties in a given network in order to provide insights into the network's characteristics [13]. In order to include the social network analysis component in the credit apportioning system, the following networks were extracted from the forum dataset (which includes all contributions for a particular offering of a course):

*Upvotes network*: An upvotes network is extracted, where an edge from user A to user B indicates that A upvoted B’s content at least once, and the weight of the edge encodes the number of times A upvoted B’s content.

*Endorsement network*: An endorsement network is extracted, where an edge from user A to user B indicates that A endorsed B’s content at least once, and the weight of the edge encodes the number of times A endorsed B’s content.

*Combined upvotes and endorsement network*: This is a union of the above two networks. An edge from user A to user B indicates that A either upvoted and/or endorsed B’s content at least once, and the weight of the edge encodes the sum of the upvotes and endorsements.

*Interaction network*: This graph models the interaction that happened on the forum over the duration of the course. In the interaction network, an edge from user A to user B indicates that B responded at least once to a question A that A posted.

**Features used in assessing forum participation**

The Credit evaluation system is a weighted combination of the following features:

**Quantitative measures**: These measures take into account the volume of contributions made by an individual. Hypothetically, if all the contributions are assumed to be of identical quality, the time and effort invested by a student is directly proportional to the volume of contributions made by him/her. Hence, the forum participation score should be directly proportional to the number of forum contributions made by the student. These quantitative measures are split into three different features in order to allow configuring varying weightage to each of them. These three features are:

- **Number of answers by the student that were endorsed**
- **Number of questions asked**
- **Number of questions answered**
- **Total number of contributions**

**Engagement level measures**: These measures are used to reward the passive engagement of the students, given that not everyone in the class might be comfortable actively posting on the forum. Some of the students are great listeners; they view or follow most of the posts and are regularly online on the forum. The two features used to apportion credit for the passive interaction are: number of days a student was online on the forum and number of posts viewed by the student.

**Qualitative measures**: These measures are used to reward the students based on the quality of their contributions. While one can readily gather the upvotes and endorsement measures from the forum dataset, these measures represent a straightforward way to add crowd sourcing to the task. High numbers indicate that the student made meaningful contributions to the forum. The use of qualitative measures also prevents the students from cheating and flooding the forum with meaningless threads (to increase their contribution count). The two qualitative features are: number of answers by the student that were endorsed, and total number of endorsements (including upvotes on the questions, answers and instructors’ endorsements) received by the student.

**Social network analysis**: In addition to the above measures social network analysis offers some less intuitive, but potentially useful measures. We use the networks described in the previous section to derive the following two features:

- **Degree Centrality in the interaction network**: Using the interaction network described in the previous section, we calculate the degree centrality for every node (or student) in the network. Degree centrality measures the number of links incident upon a node [15]. Higher degree centrality of a student implies that the student answered questions or resolved doubts for a large number of students. On a high level, degree centrality in the interaction network translates to the "helpfulness" of the student.

- **Page rank in the combined network**: The combined upvotes and endorsement network was used in order to capture importance in both upvotes and endorsement information using a single metric. Page rank can additionally help uncover influential or important students in the network [11]. Their reach extends beyond their immediate neighbors, and is therefore not captured by the earlier described upvote/endorsement figures. The higher the page rank in the combined network, the more central the student.

To obtain the ground truth for training the system, human experts were consulted once again. Given the high course enrollment of 700+, student contributions were again sampled. The next section discusses how the authors selected students whose contributions were to be included in the experts survey, and the procedure followed for ground truth creation.

**Selection of survey items and collection of ground truth**

This time each item in the survey for the experts was a pair of two posts by different students. The experts were asked to
We then sorted these ten rank differences in descending order. In order to avoid fatiguing the experts, the set was again partitioned into two batches such that each question or answer pair was voted on by at least 12 experts. The 24 experts were students who had taken the same course or a related course from which the forum contributions were drawn.

The survey instructions were as follows:

Each of the following sections presents one pair of questions or answers that were posted to the course forum in the past. We ask that you indicate for each pair the contribution that might have been most helpful to the rest of the class.

One sample item from the survey is as follows:

Q1: I am very confused about alpha-beta pruning, as we do not have example code from lecture. When we say we prune certain leaf, what does it mean? Does it mean we do not store that choices?

Q2: To create our own label, must it been binary label 1,-1? or it can be multi-categories with labels of any number? Is the feature still word counts or can be anything?

Which of the above two questions contributes more to the class community?

In order to learn the experts’ intuition about which predictors might be important in ranking student forum contributions, the following related question was introduced in the ranking survey. The question was presented to the survey takers once, at a random time in their processing the survey, such that they could drag the entries up and down to arrange the predictors in decreasing order of relevance.

Imagine you had the following statistics about forum contributions by all students at the end of the term. In your opinion, which statistics are important to evaluate the forum contributions to the class. Please drag the entries up and down to indicate their relative importance. The first entry would be the most important.

- Number of questions asked by the student
- Number of questions answered by the student
- Total number of posts viewed by the student
- Total number of days the student was online on the forum
- Total number of endorsements received by the student
- Total number of Forum contributions by the student (including questions, answers, notes, follow-ups, etc.)

Based on the majority vote for every rank, we arrive at a ranking order using the experts’ intuition. The results are summarized in Table 3. Rank 1 is the most important feature. The percentage of experts agreeing with each ranking is also included.

**Experiment: Forum Participation Credit**

A broad choice of machine learning approaches is available for ranking. This abundance is due to the relevance of ranking to traditional information retrieval tasks [9] and Web search [2]. Comparisons of these methods are available in the literature (e.g. [4]).

The pairwise comparison approach we used for our experts is mirrored in a number of widely used methods for having
Table 3: Relative ordering of the features based on majority expert opinion and the percentage support for each ranking

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>% support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td># of endorsements</td>
<td>60.7</td>
</tr>
<tr>
<td>2</td>
<td># of questions answered</td>
<td>57.1</td>
</tr>
<tr>
<td>3</td>
<td># of Forum contributions</td>
<td>46.4</td>
</tr>
<tr>
<td>4</td>
<td># of questions asked</td>
<td>46.4</td>
</tr>
<tr>
<td>5</td>
<td># of posts viewed</td>
<td>60.7</td>
</tr>
<tr>
<td>6</td>
<td># of days online</td>
<td>64.2</td>
</tr>
</tbody>
</table>

machines learn to rank. We chose the common support vector machine (SVM) for the task [6] [8]. This mechanism learns a function that separates groups of items. In this case the items are forum contributions, and the separation criterion is the degree of usefulness for the class.

Our training set consisted of 68 expert-ranked pairs of forum posts with the values of our predictors added. We held out 21 of our expert rankings as a test set. For our implementation we used SVM$^\text{rank}$ [8].

**Results: Forum Participation Credit**

The SVM attempts to minimize the number of item pairs that must be flipped in order to turn the machine’s prediction into the ranking that matches the experts. The number of necessary flips is thus a measure of performance.

Training with parameter $C = 67$ (equivalent to $C = 1$ in [7]) the subsequent validation correctly predicted rank order in 76% of cases. That is, of the 21 test pairs the finished SVM correctly predicted 16 pairs, while mistaking the rank order for five items.

**Discussion: Forum Participation Credit**

The work of the SVM produces a partial order of forum contributions, and therefore by proxy of the students who contributed those posts. We examined the 42 rankings of the test set to understand how the algorithm would impact credit distribution for this small set.

We observed that the 42 contributions were not a full ordering. Instead, as shown in the histogram of Figure 5, equivalence classes of contribution quality emerged. The distribution was bimodal rather than normal, which is probably to be expected.

**Implementation Design**

In order to incorporate the above findings into a real world usable system, additional considerations need to be made.

**Spam detection**

Since the forum has a collaborative students’ answer section, there can be instances when students try to game the system by making a minimal edit to someone else’s answer (e.g. insert a punctuation or a space) to increase their answers count. To counter this behavior, the system generates the list of all the students who have edit distances less than a threshold, which is configurable by instructors. By default, this threshold edit distance is set to 15. Thus, the system helps flag students who try to cheat, and allows the instructors to factor such attempts into their grading.

**Distribution of Rank Induced Groups**

Figure 5: Distribution of rank induced forum contribution rankings.

**Customization by instructors**

In order to allow the instructors to retain complete autonomy over their grading scheme, each of the parameters in the system is made configurable by the instructors. Thus, the instructors can tune the system to what works best for their class. For example, if an instructor believes that for his particular class, answering other students’ questions is the most important indicator of high quality forum participation, then the weightage of number of questions answered can be configured to be the highest.

**WEAKNESSES AND FUTURE WORK**

We presented an evolving computing system around the online forum facilities that are common in distance learning, and are increasingly deployed in residential settings as well.

Our focus was on the automation of two tasks that will boost the value of forum use over time. The first task we tackled was the automated construction of frequently asked questions from forum posts. We asked experts for judgments over samples of posts. These judgments served as ground truth for the training of a logistic regression classifier.

The automation of a second task, the principled apportioning of academic credit, again involved experts for the creation of ground truth. In this case we trained a support vector machine to learn how forum statistics can be used to approximate the expert judgments.

All machine learning approaches benefit from large amounts of training data. Yet the use of human expert labor to produce those data is expensive. The creation of neither gold sets can be accomplished by, for instance, a Mechanical Turk job. Staff familiar with the graduate level course were required. Both our experiments would benefit from additional training data. Apart from quantity, the small training sets contained some
class imbalances that were unavoidable. Many classifiers train best from data that include roughly equal numbers of examples from all classes.

Nevertheless, the trained models’ respectable performance measures are significant. The accomplishments of the algorithms must be compared with current practice. Today’s scenarios are dominated by the large post volumes that severely limit the (maybe) optimal use of human judgment over the entire forum.

Going forward we plan to feed the result of the academic credit apportioning task back into the FAQ list construction effort. We further plan to explore how much the addition of content analysis can boost our current results.

Another line of follow-on research needs to be an investigation into the generalizability of our results to non-science classes. The success of such transfer learning is not guaranteed, but is not out of the question.

We propose that the records from online course forum activity are important investments of human cognition. That investment is expensive, and it is to every stakeholder’s advantage to maximize the return on that effort. The work presented here is a step towards such maximization.

REFERENCES