# ASSESSING STUDENT LEARNING WITH AUTOMATED TEXT PROCESSING TECHNIQUES

*Yi-fang Brook Wu*
Department of Information Systems
College of Computing Sciences
New Jersey Institute of Technology

*Xin Chen*
Department of Information Systems
College of Computing Sciences
New Jersey Institute of Technology

## ABSTRACT

Research on distance learning and computer-aided grading has been developed in parallel. Little work has been done in the past to join the two areas to solve the problem of automated learning assessment in virtual classrooms. This paper presents a model for learning assessment using an automated text processing technique to analyze class messages with an emphasis on course topics produced in an online class. It is suggested that students should be evaluated on many dimensions, including the learning artifacts such as course work submitted and class participation. Taking all these grading criteria into consideration, we design a model which combines three grading factors: the quality of course work, the quantity of efforts, and the activeness of participation, for evaluating the performance of students in the class. These three main items are measured on the basis of keyword contribution, message length, and message count, and a score is derived from the class messages to evaluate students' performance. An assessment model is then constructed from these three measures to compute a performance indicator score for each student. The experiment shows that there is a high correlation between the performance indicator scores and the actual grades assigned by instructors. The rank orders of students by performance indicator scores and by the actual grades are highly correlated as well. Evidence from the experiment shows that the computer grader can be a great supplementary teaching and grading tool for distance learning instructors.

## KEYWORDS

Virtual Classroom, Distance Learning, Learning Assessment, Course Messages, Text Processing, Noun Phrase

## I. INTRODUCTION

Advances in both theory and technology of "Virtual Classroom[s]" [1] have contributed to the popularity of distance learning courses delivered online. Many traditional face-to-face classes also take advantage of this delivery channel by having an online electronic conferencing system that supports class communication after in-class meetings. In such courses, instructors encourage students to discuss course topics for knowledge sharing online, and they may request students to submit assignments to the system for easy course management. In an active class, a large number of text messages and attachments could be generated in a single semester. Students' work submitted online usually accounts for a large portion of their final grades. To accurately assess student work, instructors have to wade through all the documents; participating in and grading the online discussion could take more than 50% of instructors' time for each online class [2]. After all those activities, remembering which student made what contribution to a discussion topic might become difficult, especially for "class participation" which requires continuous contribution from students for the duration of the course.

Furthermore, different instructors may have different grading preferences, which could lead to bias in determining students' grades. Combining the assessments from multiple judges has been proven useful for increasing the accuracy of the final decision [3]. Therefore, in addition to the instructor who solely makes judgments on student performance, a second grader is useful and beneficial. Using a second human grader, however, is not feasible in most distance learning classes because of the limit on human resources. An automated learning assessment system, thus, would be a great aid to instructors by reducing their workload and providing a second set of grades for references.

Research on virtual classrooms has reported various theoretical models and practical approaches for learning assessment. Although they are effective, these methods require additional effort from instructors (e.g. classifying class discussions manually). On the other hand, research on computer-aided grading aims at assessing the quality of students' work automatically. It has achieved an acceptable or satisfactory level of performance: the correlation between automatic grades and human judgments ranges from 0.4, to 0.9, which is comparable to the correlation between human judgments [44]. Because these approaches have been focused on assessing the quality of single objects (e.g. an essay, a piece of program, etc.), they cannot be applied directly to learning assessment of discussions and assignments to which students have contributed more than once during a certain period of time. However, as such automated grading approaches mainly deal with textual objects expressed in natural language, they could be a solution to learning assessment, if they can be re-designed to suit the characteristics of discussions and assignments in online classes. Little work has been done to bridge the gap between the two areas of learning assessment and computer-aided grading.

This paper addresses the problem of automated assessment of student learning within a period of time, and presents a novel assessment model which predicts the class performance of students using automated text processing techniques. In distance learning classes, students are expected not only to hand in high-quality solutions to assignments, but also actively to participate in online class discussions. In general, three aspects are the main factors of class performance evaluation, especially of class discussions and participation. They are: the quality of students' work, the quantity of their efforts, and the activeness of their participation. To extract such information from class messages with automated text processing techniques, we derive three measures: keyword contribution (KC), message length (ML), and message count (MC). We measure each aspect. The three measures are combined into a linear model, in which each measure accounts for a certain proportion of a final score, called a performance indicator. The experimental results show that the assessment model works well in terms of correlating with human graders.

Although the results are promising, we do not expect the computer grader to entirely replace human instructors in evaluating student class performance. However, the model can be implemented as a supplementary teaching tool, serving as a second grader and helping instructors to make better judgments on students' work. The supplementary teaching tool is also great for course management by keeping track of students' online contributions. With the tool, it is easy to show which student posted what messages in which conferences, and to compare students' contributions even without using the automatic grading function.

The remainder of this paper is organized as follows. Section two presents existing research on both learning assessment in virtual classrooms and automated essay grading. The proposed assessment model is described in detail in section three. The experimental design and the results are presented in section four. After the discussions on issues arisen from the study, we conclude the paper with future research directions.

# II. LITERATURE REVIEW

From the literature we identified two streams of research that are related to our study—learning assessment in virtual classrooms and computer-aided grading.

## A. Learning Assessment in Virtual Classrooms

Good learning assessment in virtual classrooms helps institutions investigate teaching and learning effectiveness, aiding faculty members in comparing the effectiveness of different course delivery methods in the non-traditional environment. The ultimate goal is to assess what students learn and how well they apply that knowledge in their course work. It is “not an end in itself but a vehicle for educational improvement” [4].

Learning assessment is “most effective when it reflects an understanding of learning as multidimensional, integrated, and revealed in performance over time” [4]. As a complex process, learning involves not only the outcomes (knowledge gained) but also the students’ ability to use the knowledge, and it is also related to values, attitudes, and habits of mind that affect academic success and performance beyond the classroom. Assessment should take a diverse array of forms to reflect these understandings. It can be argued that the greater the diversity is in the methods of assessment, the fairer the assessment is to students [5]. Therefore, multiple measures related to individual academic program and course objectives should be used in studying student performance [6, 7]. Some of the more commonly used assessment methods are traditional closed-book exams, open-book exams, essays, reviews, reports, and presentations. In virtual classrooms, these methods are still usable to assess what students have learned. In addition to assessing the quality of learning outcomes, evaluating the participation through which students gain the knowledge is important as well. In distance learning courses, evaluating student participation requires a more delicate method.

Collaborative learning can increase student achievement and high-level thinking [8], and student participation is a key to effective collaborative learning [9]. When students are actively involved in collaborative learning online, the outcomes can be as good as or better than those of traditional classes [10]. These research findings and observations indicate that students need to be active participants in an online course in order to succeed. Assessing student class participation has become an important part of learning evaluation. In most of the distance learning classes, instructors give a certain proportion of the final grade to online participation.

One way to identify effective participation is utilizing Bloom's *Taxonomy of Educational Objectives* [11] to interpret discourse [12]. The taxonomy identifies six objectives: knowledge, comprehension, application, analysis, synthesis and evaluation. Each discussion message is manually classified into one of the six objectives, and the distribution on all objectives reflects the learner's ability to formulate value judgments about theories and methods. Another similar approach analyzes messages from four educational dimensions—interactive, social, cognitive and meta-cognitive—as well as the frequency, structure and type of on-line participation [13]. Limitations of the above methods include the difficulty to implement with less-structured online discussions and the difficulty for assessors to make consistent judgments, and the increasing workload of instructors.

Besides the above content analysis approaches, using information on students' usage of the system, such as login times and number of posts, to evaluate participation is also explored. Along with the content analysis, such information enables more detailed and accurate interpretation of a student's participation. A common element for learning in a typical classroom environment is the social and communicative interactions between students and their teacher, and between students and students. Collaborative learning theory highlights group interaction, which is often viewed as a major learning factor in collaborative learning. There are learner-content, learner-instructor, and learner-learner interactivity [14].

Measuring interactivity in the class has been considered in evaluating success in interactive web-based teaching [15]. It can "lead to an evaluation of the levels of collaboration at work among learners, of their active participation in the accumulation of knowledge, and of their skills in structuring the information presented on-screen" [16].

## B. Computer-aided Grading

The idea of developing computer programs to grade students' work was initiated in 1960s [17]. Recently, more theoretical models and practical implementations have been proposed to grade various types of students' works, such as computer programs [18], prose [19], language tests [20], and essays [21, 22, 23, 24, 25].

Student programs can be graded with a software testing approach, in which a testing framework provides guidance for developing the assignment specification and the grading program [18]. In the WebLAS (*Web-based Language Assessment System*) [20], students' free responses to questions are automatically scored. The system learns the grading criteria when instructors and language experts create tasks. They provide the system with task input and prompts, as well as interactively inform the system how to score student responses.

Essays can also be graded by computer programs. There are two types of automated essay grading approaches: surface feature based and content based. The former is developed upon the idea that the quality of an essay could be revealed by certain surface features, which would correlate to the grades assigned by human judges. Project Essay Grade (PEG) [17] extracts linguistic features from training essays and uses a multiple regression model to develop an equation to predict the grades of new essays. The latter focuses on the semantic relationships between words and the context. A semantic space, i.e. the contextual usage of words, is constructed from training essays. A test essay is then compared with the documents in the space, and assigned a score according to the grades of the nearest essay(s). Early work in Educational Testing Services (ETS) [26] achieves essay grading by correctly classifying the answers by content at sentence level. On the other hand, Latent Semantic Analysis (LSA) model [23] discards all linguistic and structure features, and operates solely on the content of essays. Training essays are converted to document-term matrices, which are then decomposed by using a Singular Value Decomposition (SVD) technique. A to-be-graded essay is converted to a vector of words and compared with all decomposed document vectors derived from training essays. The score of the most similar training essay is assigned to the to-be-graded essay as its grade.

A hybrid approach utilizes both types of features. More recent work at ETS [21] has focused on a "Hybrid Feature Technology" for essay grading. The system, E-rater, takes both linguistic and content features into consideration. Larkey [24] applied text categorization techniques to classify essays based on content. Different classifiers are trained to classify essays into appropriate categories, and grades are computed according to the marks of the neighbor(s) in the same category. The content-based score, along with 11 text features, is entered into a multiple regression model to predict the grades of new essays.

## C. The Gap

Although automated grading approaches are effective in achieving their goals, they are not suitable for assessing classroom learning that accumulates over a period of time. The reasons include:

- Essay grading approaches aim at assessing the quality of a single essay rather than a set of messages accumulated over one semester. Even if the quality of each individual message could be correctly estimated, the sum might not reflect the student's actual performance, because the content of the messages are not independent.
- Some approaches take into account writing styles, which are not so important in grading online discussions that consist of many informal messages.
- Most of the approaches do not perform well with short texts (less than 100 words). However, short messages are very common in online class discussions.
- All existing approaches require a large set of training data to teach the computer system the grading criteria. However, human rated training class messages for a whole semester are expensive and nearly impossible to obtain.
- These approaches do not consider other factors in grading, such as frequency of participation and interaction with other participants.

Automated grading draws on computing and language models, while neglecting other factors that are addressed in learning assessment research. Although research in both streams discussed above has been well developed, little work has been done to join them together to solve the particular problem—automated evaluation of student participation in distance learning classes. This study attempts to tackle the problem with automated text processing techniques.

# III. THE ASSESSMENT MODEL

The goal of this study is to access student learning by analyzing the class messages produced by students and instructors in the e-learning system. The scope of our study focuses especially on assignments that require student participation over a period of time, e.g. directed discussions with assigned topics or undirected discussions without assigned topics. In these kinds of class assignments, assignment submissions and discussion messages are expected to occur more than just once. The instructor might easily lose track of students' activities because of large amounts of messages posted. On the other hand, one-time-only types of assignments, such as final papers, are not suitable for analysis by our tools. This section presents the assessment model, including the basic concepts concerning keyword contribution, message length, and message count, as well as the assessment model consisting of the three measures.

The model assesses student learning from three aspects: the quality of their course work, the quantity of their efforts, and the activeness of their participation. Three measures—keyword contribution, message length, and message count—are derived from the class messages to measure each assessment aspect respectively.

## A. Keyword Contribution Mining

Keyword contribution uses content analysis of messages to measure the quality of a student's work. Below we give the basic definitions of the concepts used throughout of the paper.

### 1. Keyword

We assume that quality of learning is revealed by the quality of messages generated by a student. The number of key concepts appearing in the messages reflects the knowledge range of the author, so the usage of key concepts could be an indicator for the learning quality.

Evidence from the language learning of children [27] and discourse analysis theories, such as Discourse Representation Theory, [28] show that the primary concepts in text are carried by noun phrases. Therefore, noun phrases are considered the conceptual entities in text messages. We define keyword as a simple, non-recursive noun phrase, i.e. a base noun phrase. A base noun phrase consists of a head and none or more modifiers, which can be adjectives or nouns.

Identifying base noun phrases from free text usually involves two sub-problems: part-of-speech (POS) tagging and noun phrase identification. The first step is more substantial, because the second step and the final result are highly dependent on the accuracy of the POS tags. A POS tagger can be supervised or unsupervised [29, 30, 31]. Supervised taggers typically rely on pre-tagged corpora to serve as the basis for the tagging process, while unsupervised taggers do not require a pre-tagged corpus but instead use sophisticated computational methods to automatically calculate the probabilistic information needed by stochastic taggers [32, 33, 34, 35] or to induce the context rules needed by rule-based systems [36, 37].

We implement a noun phrase extractor by using a lexical database to estimate the initial lexical probability of each word and then disambiguating a multi-tag word by examining its previous *n* (2~4) tokens against a list of manually defined syntactic rules. The free text is first tokenized. A simplified WordNet database [38], which contains words divided into four categories (noun, verb, adjective, and adverb) and the number of senses of each word in each of the categories, is used to assign the initial POS tag, which is determined by selecting the category with the maximum number of senses. If a word is found in more than one category, it is marked as a multi-tag word.

The second stage is multi-tag disambiguation. For each multi-tag word, the sequence of the POS tags of the previous *n* tokens is examined against a list of predefined syntactic rules. For example, "*hit*" can be either a noun or a verb. If the previous word is a determiner (*the, a, this*, etc), it will be tagged as a noun rather than a verb, and the multi-tag mark is removed. If none of the rules is matched, some heuristics are used. For instance, if a word is found in both the noun and the verb category, but ends with "*tion,*" it is tagged as a noun.

After tagging the text, the noun phrase extractor identifies noun phrases by selecting the sequence of POS tags that are of interest. The current sequence pattern is defined as [A|N] N, where A refers to Adjective and N refers to Noun. The pattern defines a base noun phrase that consists of a head N and none or more modifiers [A|N].

### 2. Class Concept Base

The unique noun phrases extracted from all class messages are defined as the class concept base, which represents the concepts related to the major topics in the class. Because the concepts in the class concept base are contributed by individual students, we can estimate a student's contribution by calculating the number of concepts contributed by the student in the class concept base. In previous studies [39, 40], keywords were treated as being equally important. However, after closely examining the extracted keywords, we found they are not equally important in terms of how significant they are in the class concept base. A weighting scheme is required to assign different weights to keywords to reflect their importance.

## 3. Keyword Weighting Scheme

We borrow the idea of term weighting in information retrieval [41] to assign weights to keywords. The importance of a keyword is measured by its frequency. The more frequently a keyword appears in a message, the more important it is with respect to that particular message. However, if a keyword is used by more students and appears in many messages, it becomes less important in terms of differentiating one student's contribution from others. In other words, a student fully contributes to the class concept base only by adding new keywords that are not used by any other students. The contribution of adding a keyword to the concept base decreases when the number of its author(s) increases. The extreme situation is that when a keyword is used by all other students, adding it to the concept base results in no contribution at all.

From another point of view, concepts contributed by more students tend to be more general. For instance, in an Information Systems course, *information systems* is likely to be used by most of the students, while *expert systems* is likely to be used by only a few of them. Messages containing too many general concepts tend to be superficial, and lack of deep analysis and strong arguments. In class discussion and other course work, we encourage students to synthesize what they have learned and add in their own understanding of the course materials. Although the usage of more specific keywords does not necessarily result in high quality work, it is still preferred because it indicates that the student is bringing in new concepts, not just repeating the existing ones.

The length of a noun phrase should also be taken into consideration. Longer noun phrases tend to be more descriptive than shorter ones. As we can see from a real example in the experiment, *COCOMO software development model* is more descriptive than *development model*, so we assign higher weights to longer phrases.

Based on the analysis above, we use the following formula to calculate weights of keywords:

$$w = \left(1 + \log(len)\right) \cdot \left( f \cdot \log \frac{N}{n} \right)$$

Above, $w$ is the weight of a keyword, *len* is the length (number of words) of the keyword, $f$ is the frequency of the keyword in the concept base, $N$ is the total number of students in the class, and $n$ is the number of students who use the keyword in their messages. We use a log function of the length to prevent it from becoming dominant when it increases. This function is similar to the tf.idf (term frequency - inverse document frequency) measure in Information Retrieval [42], but the inverse frequency of a keyword in our study is across students, not messages.

## 4. Keyword Contribution

After assigning weights to keywords, we calculate Keyword Contribution (KC) by adding up the weights of the unique keywords contributed by each student and dividing the result by the sum of weights of all unique keywords in the concept base. It is denoted as follows:

$$KC_i = \frac{W_i}{W}$$

where $KC_i$ is the keyword contribution for student $I$; $W_i$ is the sum of the weights of the unique keywords contributed by student $I$; and $W$ is the sum of the weights of all unique keywords in the class concept base.

KC measures the quality of a student’s messages by calculating how many of the class concepts are contributed by the student in all his/her messages. We aggregate all messages from each student, because the learning process is accumulative and the knowledge inside each message is not independent. In other words, using a keyword a student already knows (appearing in his/her previous messages) in a new message does not necessarily mean that the student learns/contributes a new concept. By aggregating all keywords and using only unique ones, we can capture the contribution of each student for the whole learning process.

In addition to the quality of a student’s work, other factors such as the efforts a student devotes to the class and the student’s activeness of class participation are considered as well. We propose two other measures for these factors—message length and message count.

## B. Message Length

Previous researches have found a direct and positive relationship between the amount of time students spend reading postings and engaged in virtual dialogue with their classmates and their achievement of course objectives [43]. Students’ effort in the virtual dialogue could be reflected by the amount of words they post to the system. Therefore, the Message Length (ML) measure is defined to measure a student’s effort in the class. ML is calculated by counting all the words (not noun phrases), no matter duplicated or not, in the student’s messages. The result is normalized by the class message size, which is the number of words in the entire class messages. Let $ML_i$ be the message length for student $i$, and $n_{ij}$ be the number of words in message $j$ of student $i$, and we have:

$$ML_i = \frac{\sum_j n_{ij}}{\sum_i \sum_j n_{ij}}$$

It is normalized to show the relative effort contributed by a student.

## C. Message Count

Activeness of participation could be measured by the logon times, but this information is not as useful as the measure we are proposing. From previous studies, it is found that “student postings constituted one indicator for actual participation in the course since it showed the number of times students read and responded in writing to the instructor's or to another student's posting” [6]. If we consider posting a message as one class activity, activeness of participation can be measured by Message Count (MC), which is the number of messages posted by a student. MC is defined as

$$MC_i = \frac{n_i}{\sum n_i}$$

## D. The Assessment Model

The three measures are combined to compute a Performance Indicator (PI) score, which is defined as

$$PI_i = \alpha KC_i + \beta ML_i + \gamma MC_i$$

Above, $PI_i$ is the performance indicator score assigned to student *i*, and the coefficients $\alpha$, $\beta$, and $\gamma$ are the weights of each of the three measures respectively. The coefficients are adjustable. Instructors can define the values by specifying the importance of each evaluation aspect. For example, by defining $\alpha$=5, $\beta$=3, and $\gamma$=1, the instructor would like to give higher grades to students who are more capable of synthesizing knowledge learned from the class, rather than posting many short content-poor messages.

# IV. EXPERIMENT AND RESULTS

## A. Experiment Design

Five classes were selected for model validation. All classes were supported by an electronic conferencing system that enabled class participants (instructors and students) to communicate by posting text message asynchronously. For each class in the electronic conference system, the instructor was able to create multiple discussion boards, so called *conferences*, to organize students' postings and discussions for different purposes, e.g. *Self-introduction*, *Assignments*, *Weekly Discussions,* and so on. The classes were chosen from different domains, and they had different designs of conference structures in the electronic conferencing system. The class information and their conference design are summarized in Table 1.

**Table 1: Summary of Class Information**

| ID | Domain | Conference Design |
|---|---|---|
| C1 | Management | Required discussions on assigned topics in the course. Students' final grades were mainly based on the points they made in each discussion. |
| C2 | Information Science | Discussions without assigned topics. Students were free to share with each other whatever topics they found relevant to the course. Grades were assigned to students' online participation, which was worth 5% of the final course grade. |
| C3 | Information Systems | Activities include required class discussion, debates, oral presentation, assignment submission, course project, and optional discussions. Some of the grading items were judged according to the documents submitted online, while others were based on hand-in materials, such as project reports, PowerPoint slides, and so on. |
| C4 | Information Systems | |
| C5 | Information Systems | |

All class messages were downloaded from the conferencing system and were converted to plain text files, from which keywords were extracted and weighed. When calculating the performance indicator score, we set the three coefficients, $\alpha$, $\beta$, and $\gamma$, to 1, because we did not know the instructors' grading preferences. However, when the grading preferences are known, it is easy to adjust the coefficients to reflect the grading preferences.

We evaluated the accuracy of the system output by comparing it with the actual grades assigned by the instructor of each class. Correlations between both the raw scores and the rank orders of students were calculated. For class 2 (C2), we chose the participation grades assigned by the instructor for comparison, because the discussion conference was designed for class participation only; while for the other four classes, students' final grades (out of 100 points) were chosen because of their course conference design.

## B. Results and Analysis

From class messages of the five selected classes, we extracted all noun phrases and assigned weights to them according to the proposed formula. As an example illustrating the keyword weighting function, table 2 shows some select keywords, as well as their weights and frequency, extracted from class 1 (C1). Keyword *company* has a high frequency (212), but its weight is still 0, because it is used by all students. It means that *company* is a general and common concept in the management domain, thus using it alone in one's messages does not bring in new concepts. Therefore, it will not contribute much to the class concept set. Given the same frequency (*f*) and the number of authors (*n*), longer phrases have higher weights (*business structure change* vs. *business decision*). Also, more frequent phrases (e.g. *business complexity* and *content management*) were assigned higher weights, and when fewer people use it, a phrase has even higher weight (e.g. *content management*). In general, the results show that the weighting function favors concepts that are longer but used by fewer authors. This is desired, because the appearance of such concepts in the messages suggests that the student focuses on domain-specific subjects relevant to certain course topics.

**Table 2: Selected Keywords and Their Weights from C1**

| Keyword | W | f | n | N |
|---|---|---|---|---|
| Company | 0.00 | 212 | 31 | 31 |
| business decision | 5.81 | 1 | 1 | 31 |
| business structure change | 7.21 | 1 | 1 | 31 |
| business complexity | 11.86 | 3 | 3 | 31 |
| content management | 17.44 | 3 | 1 | 31 |

w: keyword weight, f: frequency, n: number of authors, N: number of students in the class

**Table 3: Summary of the PI scores**

| | Range | N | Mean (Std. Dev.) |
|---|---|---|---|
| C1 | 0.01~0.23 | 31 | 0.13 (0.06) |
| C2 | 0.01~0.53 | 27 | 0.16 (0.15) |
| C3 | 0.16~0.51 | 15 | 0.28 (0.11) |
| C4 | 0.04~0.38 | 17 | 0.18 (0.10) |
| C5 | 0.00~0.47 | 19 | 0.23 (0.12) |

The performance indicator score of each student was calculated using the three measures derived from the class messages. Table 3 summarizes the PI scores of the five classes. To examine how accurately the model assesses students' class performance, we compared the PI scores with the students' actual grades assigned by instructors. The Pearson product-moment correlations between the PI scores and the actual grades were calculated. Correlations between the individual measures and the actual grades were also calculated. The results in the second column of Table 4 ($r_{PI\text{-}G}$) demonstrate that there is a high correlation between the PI scores and the actual grades (from 0.62 to 0.92). According to a report in the essay grading literature, agreement between computer graders and human judges varies from 0.4 to 0.9 approximately, and that is comparable to or even better than agreement between two human graders [44]. Although we could not compare our results directly to the agreement between two instructors who teach the same class independently, according to the results in essay grading literature, it is reasonable to conclude that our model performs well in terms of correlating with human evaluators.

We also calculated the correlations between the three measures and the actual grades. They are shown in Table 4 as $r_{KC\text{-}G}$ , $r_{ML\text{-}G}$ and $r_{MC\text{-}G}$ which stand for the correlations between the actual grades and *KC*, *ML*, *MC* respectively. The results show that, in most cases, *KC* performs slightly better than *ML* and *MC*, and *PI* performs better than any of the three measures.

**Table 4: Correlations**

| | $r_{PI\text{-}G}$ | $r_{KC\text{-}G}$ | $r_{ML\text{-}G}$ | $r_{MC\text{-}G}$ | $r_{ro}$ |
|---|---|---|---|---|---|
| C1 | 0.92 | 0.90 | 0.87 | 0.88 | 0.94 |
| C2 | 0.87 | 0.85 | 0.84 | 0.84 | 0.98 |
| C3 | 0.62 | 0.54 | 0.58 | 0.77 | 0.74 |
| C4* | 0.80 | 0.79 | 0.78 | 0.74 | 0.94 |
| C5 | 0.62 | 0.65 | 0.50 | 0.52 | 0.63 |

$r_{PI\text{-}G}$: Correlation between the PI scores and the actual grades
$r_{KC\text{-}G}$: Correlation between the KC scores and the actual grades
$r_{ML\text{-}G}$: Correlation between the ML scores and the actual grades
$r_{MC\text{-}G}$: Correlation between the MC scores and the actual grades
$r_{ro}$: Correlation between the rank orders (*Rg* and *Rpi*) of students
*: Calculation is after the removal of an outlier

Unlike essay grading approaches which attempt to assign each essay a concrete score, our model does not attempt to predict the actual grades of students. It assesses students by comparing them with others and distinguishing “strong” students from “weak” ones. The PI scores could help instructors assess students by grouping them at different levels, which is still a common approach used in practice. Therefore, the rank order of students by the *PI* scores would be more interesting to instructors. To evaluate how well the PI scores rank students, we computed the correlations between the rank orders of students by the PI scores and by the actual grades. First, students are ranked by their actual grades in descending order, and the rank order is recorded as *Rg*. Similarly, another rank order, *Rpi*, ranks students by their *PI* scores. The Spearman ranker order correlation between *Rg* and *Rpi* is then calculated. The result is shown in the last column of Table 4. The high correlation between *Rg* and *Rpi* suggests that the *PI* scores rank students correctly.

## V. DISCUSSION

Unlike many essay grading approaches which analyze writing styles and grammar, our model does not take those factors into consideration. Writing styles are important for composing high quality essays, because a good writer, most likely, will avoid using the same keywords to convey the same concepts. However, in virtual classrooms (except online writing classes), instructors are more concerned with what is expressed in the work of students than how the messages are composed. In addition, each student is graded based upon his/her work in the entire learning process (e.g. one semester), not just a single essay. The majority of the work, such as discussions and debates, is often in an informal format. Therefore, writing style measures for essay grading may not apply to assessing the quality of a set of class messages. However, for assignments that are required to be answered in essay format, it is possible to extend our model by applying an essay grading approach to them and combining the results with the current model.

In the experiment, we found that the performance indicator scores generated from the model were highly correlated with the actual grades assigned by instructors. Although the correlation between the judgments of two instructors grading the same class independently is unknown, it is reasonable to assume that such correlation is comparable to what has been reported in the automatic essay grading literature, which is 0.73 in [22] and 0.69 in [23]. The experiment results, thus, suggest that the performance of our model is comparable to, if not better than, that of a human instructor. The validity of our model, therefore, is established. However, we do not expect the computer grader to entirely replace the instructor and solely judge the performance of students in an online class. The model can be implemented as a teaching tool to help instructors obtain a reference to students' performance without wading through the huge amount of class messages, which is a tedious and intensive procedure when performed without the use of automated text processing techniques. The tool could be employed as a supplementary grader to help instructors make better judgments with reduced workload.

Evidence of the supplementary role of the computer grader is found in the experiment. PI scores deviated from the actual grades may suggest either inappropriate grades, or something special in the messages, or both. In C2, the PI score of one student is relatively higher than the actual grade. By reexamining the student's messages, the instructor found that the student copied and pasted a long message along with the source URL from the web without adding his/her personal opinions. Even though the instructor had encouraged students to share anything they found relevant and interesting to the class, without personal opinions and thoughts, the instructor considered copying and pasting a lesser effort. Therefore, the original grade is confirmed.

A similar case was found in C4, in which a student (hereafter known as S) got the highest PI score, but a low grade. The instructor explained that S was an exception in that class as S submitted almost every assignment late because of family and personal medical problems. The instructor accepted S's late assignments but gave low grades. After the makeup exam, the instructor changed S's final grade to a higher one. For this reason, we excluded S from analysis when calculating the correlations (see Table 4 for details). Figure 1 illustrates the close relationship between *Rg* and *Rpi* of C4 except the outlier S discussed above.

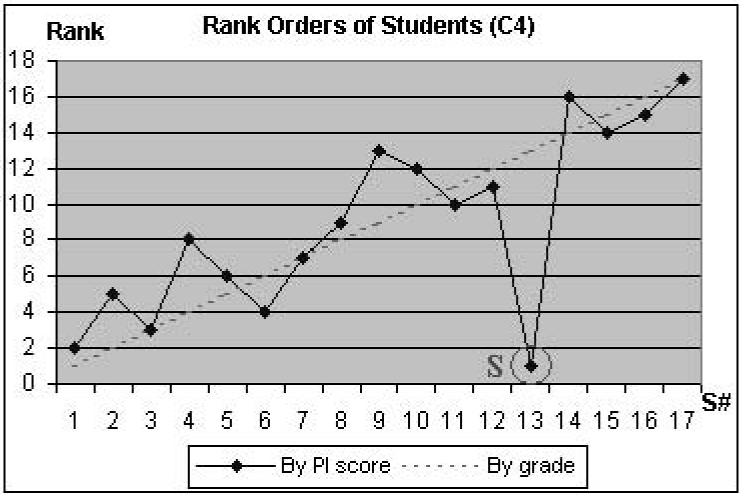

**Figure 1: Rank Orders of Students by PI Score and by Grade (C4)**

Having the program serving as a second grader, instructors are able to capture outliers, and to reduce misjudgment, bias, or errors in grading.

## VI. CONCLUSIONS AND FUTURE RESEARCH

We have presented a model for automated assessment of student learning in virtual classrooms. This model uses automated text processing techniques to calculate performance indicator scores for students to predict their class performance. The experiment results show that there is a high correlation between the PI scores and the actual grades assigned by instructors, for both the absolute scores and the relative rank orders of students. We also found the evidences from the experiment of the usefulness of the tool as a supplementary teaching tool.

The correlations presented in Table 4 are generally high, but we do find C3 and C5 have relatively low correlations. By looking into the conference boards, we identify the following possible reasons:

- The conference boards have some files that cannot be processed by the current version of our program, such as attachments, PowerPoint slides, and audio files. However, they were manually evaluated by the instructors.
- There are private conferences that are not accessible to the account we used to download messages. For example, conferences created for individual group projects could only be accessed by group members and the WebBoard manager, i.e. the instructor.

The observations suggest some improvements that could be made to the program:

- The program should allow instructors to assign different weights to conferences. For example, instructors might want to give a low weight to the self-introduction conference board. Messages in a conference should inherit its weights, and the weights should be further applied to the calculation of the three assessment measures.
- The program should have the ability to process other types of files, such as Word document, PDF attachments and PowerPoint slides.

Also, it will be interesting to investigate the timestamps of messages, because they may reveal a student’s activities at a certain point during the class. Introducing a time dimension into the model may enable us to evaluate students at different time points, and to ultimately develop a measure for learning that takes place throughout the class. Another potential way to improve the model is to consider students’ interactions in the class. The current model does not differentiate initial postings from replies. The Social Network Analysis (SNA) technique [45, 46] could be used to evaluation students’ participation by analyzing the interactions among students.

## VII. ACKNOWLEDGEMENTS

The authors especially thank Professor Murray Turoff for supporting the development of the evaluation tool. We thank Professors Roxanne Hiltz and Il Im for providing us the evaluation data sets. We also thank our anonymous reviewers for their helpful comments.

## VIII. ABOUT THE AUTHORS

**Yi-fang Brook Wu** is an Assistant Professor in the Information Systems Department at New Jersey Institute of Technology. She is interested in designing systems which automatically analyze textual entities and their relationships in a large document collection.

**Xin Chen** is currently a PhD student in the Information Systems Department at New Jersey Institute of Technology. His research interests are centered on techniques for discovering knowledge from text, especially when the user's background knowledge or interests are involved. He is also interested in various text processing techniques, such as web search engine, noun phrase extracting, document classification and clustering, and their applications in business and education.

## VIII. REFERENCES

1. **Hiltz, S. R.** *The Virtual Classroom: Learning Without Limits Via Computer Networks*. Norwood NJ: Ablex, 1994.
2. **Lazarus, B. D.** Teaching Courses Online: How Much Time Does it Take? *Journal of Asynchronous Learning Networks* 7(3): September 2003.
3. **Winkler, R. L. and R. T. Clemen.** Multiple Experts vs. Multiple Methods: Combining Correlation Assessments. *Decision Analysis,* November 2002.
4. **Astin, A. W., T. W. Banta, K. P. Cross, E. El-Khawas, P. T. Ewell P. Hutchings, T. J. Marchese, K. M. McClenney, M. Mentkowski, M. A. Miller, E. T. Moran, and B. D. Wright.** 9 Principles of Good Practice for Assessing Student Learning. American Association for Higher Education, 2003. Online: http://www.aahe.org/assessment/principl.htm.
5. **Race, P.** The Art of Assessment. *SEDA publication the New Academic* 5(3): 1995.
6. **Picciano, A.** Beyond Student Perceptions: Issues Of Interaction, Presence, and Performance In An Online Course. *Journal of Asynchronous Learning Networks* 6(1): July 2002.
7. **Shea, P., E. Fredericksen, A. Pickett, W. Pelz, and K. Swan.** Measures of learning effectiveness in the SUNY Learning Network. In Bourne, J. and Moore, J. C. (eds) *Online Education,* Volume 2, Needham, MA: Sloan-C, 2001.
8. **Kitchen, D. and D. McDougall.** Collaborative learning on the Internet. *Journal of Educational Technology Systems* 27(3): 1998–99.
9. **Hardless, C., J. Lundin, and U. Nulden** Mandatory Participation in Asynchronous Learning Networks. *Proceedings of HICSS*, Maui, USA, January 2001.
10. **Hiltz, S. R., N. Coppola, N. Rotter, M. Turoff, and R. Benbunan-Fich.** Measuring the Importance of Collaborative Learning for the Effectiveness of ALN: A Multi-Measure, Multi-Method Approach. *Journal of Asynchronous Learning Networks* 4(2): 2000.
11. **Bloom, B.** *Taxonomy of educational objectives: The classification of educational goals*. Handbook I, cognitive domain. New York: Longman, 1956.
12. **Peat, M.** Online assessment: The use of web based self assessment materials to support self directed learning. *Proceedings of the 9th Annual Teaching Learning Forum*, 2–4 February 2000.
13. **McKenzie, W. and D. Murphy.** “I hope this goes somewhere”: Evaluation of an online discussion group. *Australian Journal of Educational Technology* 16(3): 239–257, 2000.
14. **Moore, M.** Three types of interaction. *The American Journal of Distance Education* 3(2): 1–6, 1992.
15. **Bodomo, A., K. K. Luke, and A. Anttila.** Evaluating Interactivity in Web-Based Learning, *Global E-Journal of Open, Flexible and Distance Education* 3(1): 2003.
16. **Henri, F.** Distance learning and computer mediated communication: Interactive quasi-interactive or monologue. In C. O'Malley (ed.) *Computer supported collaborative learning,* NATO ASI series, Berlin: Springer-Verlag, 1995.
17. **Page, E. B.** Grading essays by computer: Progress report. Notes from the 1966 Invitational Conference on Testing Problems, 87–100, 1966.
18. **Jones, E. L.** Grading student programs - a software testing approach. *The Journal of Computing in Small Colleges* 16(2): 2001.
19. **Page, E. B.** New Computer Grading of Student Prose Using Modern Concepts and Software. *The Journal of Experimental Education* 62(2): 127–142, 1994.

20. **Bachman, L. F., N. Carr, G. Kamei, M. Kim, M. J. Pan, C. Salvador, and Y. Sawaki.** A reliable approach to automatic assessment of short answer free responses. *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.
21. **Burstein, J., C. Leacock, and M. Chodorow.** CriterionSM: Online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, August 2003.
22. **Foltz, P. W., D. Laham, and T. K. Landauer.** The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* 1(2): 1999.
23. **Landauer, T. K. and J. Psotka.** Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments* 8(2): 73–86 2000.
24. **Larkey, L. S.** Automatic essay grading using text categorization techniques. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998.
25. **Page, E. B.** Computer Grading of Essays: A Different Kind of Testing? Address for APA Annual Meeting, Aug. 13, 1995.
26. **Burstein, J., R. Kaplan, S. Wolff, and C. Lu.** Using Lexical Semantic Techniques to Classify Free-Responses. *In Proceedings of SIGLEX 1996 workshop*, Annual Meeting of the Association of Computational Linguistics, University of California, Santa Cruz, 1996.
27. **Snow, C. E. and C. A. Ferguson.** *Talking to Children: Language Input and Acquisition*. Cambridge: Cambridge University Press, 1997.
28. **Kamp, H. A.** Theory of Truth and Semantic Representation. In J. Groenendijk, T. Janssen, and M. Stokhof (eds.) *Formal Methods in the Study of Language*, Vol. 1. Mathema-tische Centrum, 1981.
29. **Brill, E. and Marcus, M.** Tagging an unfamiliar text with minimal human supervision. ARPA Technical Report, 1993.
30. **Brill, E.** Unsupervised learning of disambiguation rules for part of speech tagging. *Proceedings of the ACL Third Workshop on Very Large Corpora*, 1–13. Somerset, New Jersey, 1995.
31. **Schütze, H.** Part-of-speech induction from scratch. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 251–258, 1993.
32. **Church, K. W.** A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *In Proceedings of the Second Conference on Applied Natural Language Processing*, 136–143, 1988.
33. **Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun.** A Practical Part-Of-Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.
34. **Dermatas, E. and G. Kokkinakis.** Automatic stochastic tagging of natural language texts. *Computational Linguistics* 21(2): 137–163, 1995.
35. **DeRose, S. J.** Grammatical category disambiguation by statistical optimization. *Computational Linguistics* 14(1): 31–39, 1998.
36. **Greene, B. B. and G. M. Rubin.** Automatic grammatical tagging of English. Technical Report, Brown University. Providence, RI, 1971.
37. **Brill, E.** Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 1995.
38. **Fellbaum, C. D.** *WordNet: An Electronic Lexical Database*. MIT Press: Cambridge, MA, 1998.
39. **Wu, Y-f. B. and Chen, X.** Assessing Distance Learning Students’ Performance: A Natural Language Processing Approach to Analyzing Online Class Discussion Messages. *Proceedings of ITCC*, Las Vegas, USA 2004.
40. **Chen, X. and Y-f. B. Wu.** Automated Evaluation of Students' Performance by Analyzing Online Messages. *Proceedings of IRMA*, New Orleans, USA, 2004.
41. **Salton, G. and C. Buckley.** Term weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5): 513–523, 1988.

42. **Salton, G.** *Automatic Text Processing: The Transformation Analysis, and Retrieval of Information by Computer*. Addison Wesley, 1989.
43. **Levenburg, N. M. and H. T. Major.** Motivating the Online Learner: The Effect of Frequency of Online Postings and Time Spent Online on Achievement of Learning Goals and Objectives. *International Online Conference on Teaching Online in Higher Education*, 2000.
44. **Williams, R.** Automated essay grading: An evaluation of four conceptual models. In A. Herrmann and M. M. Kulski (Eds), *Expanding Horizons in Teaching and Learning*. Proceedings of the 10th Annual Teaching Learning Forum, 7–9 February, 2001.
45. **Wasserman, S. and K. Faust.** *Social Network Analysis*. Cambridge University Press, 1994.
46. **Reffay, C. and T. Chanier.** How Social Network Analysis Can Help To Measure Cohesion in Collaborative Distance-Learning. *Proceeding of Computer Supported Collaborative Learning Conference*, June 2003.