IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4889

Automated Evaluation of Students' Performance by Analyzing Online Messages

Xin Cheng
GITC 4215, New Jersey Institute of Technology, University Heights, Neward, NJ 07102, USA, xc7@njit.edu

Yi-fang Brook Wu
GITC 4215, New Jersey Institute of Technology, University Heights, Neward, NJ 07102,USA, wu@njit.edu

ABSTRACT

Students participate in a virtual classroom and interact with the instructor and other students largely by composing text messages and replying to others. We propose measurements derived from natural language processing techniques to evaluate these text messages. Students' performance is evaluated from three perspectives: knowledge they learn from, effort they devote to, and their participation activeness in the class; three measures - keyword density, message length, and message count, are derived for each evaluation aspect respectively. An overall performance indicator is computed from the three measures. The experiment shows that there is a high correlation between the performance indicator scores and the actual grades. The rank order of students by the performance indicator score and that by the actual grades are highly correlated as well.

INTRODUCTION

The advances in information technology and theories in Asynchronous Learning Network (ALN) (Hiltz, 1994) have populated the online courses. A large number of students complete part of, or even all their courses through this electronic channel. Many traditional face-to-face courses also take advantage of the convenience and efficiency of the asynchronous teaching and learning mode by having an online discussion tool for communication after class meetings. In such conditions, participants communicate with each other largely by composing text messages and replying to others. Because of the large volume of the text messages, it is very costly for instructors to read them to grade students. An automated performance evaluating system would be a great aid to instructors.

The resistance to using a computer program as the only judge of students' work is well justified, so any automated computer grader should not be supposed to replace the human graders. The computer grader, however, could serve as a second grader, and supplement to the judgments of the instructor, who is usually the sole evaluator in the class. It could help instructors improve their grading by enabling them to reevaluate students when disagreements occur.

Using computer programs to grade students' work has been a long-time interest to researchers. The idea was initiated back to 1960s (Page, 1966). Recently a few more theoretical models, as well as practical implementations, have been reported. The results are encouraging, with the correlation between the computer grader and the human judges as good as the one between the human graders. However, majority of the work has been focusing on assessing the quality of individual essays; it cannot be applied directly to grading students' online work, which is accumulative over a semester.

In this paper, we are interested in exploring a new and simple method to computer evaluating of students' performance. It measures a student's performance from three perspectives: (1) how well, (2) how much, and (3) how often the student contributes to the class message set. Three features of the online text messages - Keyword Density (KD), Message Length (ML), and Message Count (MC), are derived for each of the perspectives respectively. A linear model is constructed to calculate the Performance Indicator (PI) score of each student. The experimental results suggest that the computer grader highly agrees with the human graders.

The following part of this paper briefly presents related work in computer aided grading, illustrates our conceptual model in detail, presents the experimental results, and concludes with our discussion and future work.

COMPUTER AIDED GRADING

Computer aided grading has been used widely in assessing various students' works, such as computer programs (Jones, 2001), prose (Page, 1994), language tests (Bachman et al, 2002), essays (Page, 1995; Larkey, 1998; Foltz, 1999; Landauer, 2000; and Burstein et al, 2003). Little work has been reported on grading students' class performance in virtual classrooms, therefore, it is worthwhile exploring the possibility of computer aided grading in this area. In general, essay grading has much in common with class performance grading, since the objects to be dealt with are both in text format and are expressed in natural language.

So far, four main streams of computer essay grading have been proposed in the literature.

The earliest model, Project Essay Grade (PEG) (Page et al, 1966), extracts linguistic features (known as *proxes*) from essays assessed by human judges and uses a multiple regression model to develop an equation, which is then used to predict the grades of new essays. Similarly, the second model, E-rater (Burstein et al, 2003), exploits linguistic features but also document structure features. A statistical model is built to relate these features to overall writing quality. Another approach, Latent Semantic Analysis (LSA) model (Landauer et al, 2000), discards all linguistic and structure features, and operates solely on the content ("bag of words") of essays. Training documents are converted into document-term matrixes (known as "semantic space"), which are then decomposed by using the Singular Value Decomposition (SVD) technique. An essay to be graded is converted to a vector of words and compared to all document vectors in the semantic space. The score of the most similar document is assigned to the essay as its grade. Text categorization techniques are used for automated essay grading (Larkey, 1998). Different classifiers are trained to assign scores to training essays. The output, along with the text features, are given to a regression model, and the equation is used to grade new essays.

Though they are proved to be effective in essay grading, these approaches are not suitable for our purpose because of the following reasons.

This conference paper appears in the book, Innovations Through Information Technology, edited by Mehdi Khosrow-Pour. Copyright © 2004, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

- Some of the grading approaches take into consideration the writing styles, which are not so important in our case. These approaches do not perform well with short texts (less than 100 words), while short messages are very common in class discussion.
- Essay grading focuses on evaluating the quality of single essays rather than the messages a student produces during the entire learning process. Even if the quality of each message could be correctly estimated, the sum might not reflect the student's actual performance, because the content of the messages are not independent.
- All existing approaches require a large set of training data to teach the computer system the grading criterion. The training data could be manually graded essays, or standard materials (e.g. chapters in a textbook). But human rated training class messages are expensive and nearly impossible to obtain in our study.
- Most of the existing approaches attempt to assign each essay with an absolute grade, while in a class, professors are more likely to grade the students relatively, e.g. by their rank orders.

This paper attempts to explore a new model suitable for evaluating students' online performance by analyzing the class text.

THE MODEL

We adopt a hybrid approach by taking account of both content and text features of the class messages. According to our experience in teaching online courses, we propose a grading scheme which evaluates students from 3 aspects: (1) the quality of their class work, (2) the quantity of their work, and (3) the activeness of their participation.

Appropriate measures for each of the evaluation aspects are derived from the text messages. They are Keyword Density (KD), Message Length (ML) and Message Count (MN).

Keyword Density

The number of key concepts a student uses reflects his/her knowledge about the topics in the course. However, it is difficult to define a fixed set of key concepts for a course, because there are always changes and updates of the course materials, and different instructors may emphasize on different aspects too. Therefore, using a fixed set of key concepts is not feasible. We assume that the key concepts covered in the current course messages, from both the instructor and the students, could be seen as a concept space for the course. By comparing the number of key concepts generated by a student to this class concept space, we could estimate how well the student does in terms of learning the existing key concepts.

The evidences from language learning of children (Snow and Ferguson, 1997) and discourse analysis theories, e.g. Discourse Representation Theory, (Kamp, 1981), show that the primary concepts in text are carried by noun phrases, which, therefore, can be considered the conceptual entities in text messages. Keyword is defined as a simple, non-recursive noun phrase. Keyword Density (KD) of a student is the proportion of noun phrases that appear in the messages generated by the student. Note that an already-appeared concept term does not expand the student's knowledge, so duplicated noun phrases by the same student are counted only once. It is denoted as

$$KD_i = \frac{NNP_i}{\sum NNP_j} = \frac{NNP_i}{NNP} \quad \cdots\cdots (1),$$

Where KD_i is the keyword density for student *i*, NNP_i is the number of distinct noun phrases in student *i*'s messages, NNP_j is the number of distinct noun phrases in student *j*'s messages, and NNP is the number of distinct noun phrases in all class messages.

Message Length

The quantity of a student's work is measured by the length of his/her messages (ML), which is calculated by counting all the individual words in the student's messages. Again, the absolute number is proportioned by the total size, the number of words in the entire class message set. Let ML_i be the message length for student *i*, NW_{ij} be the number of words in message *j* of student *i*, and NW_{kj} is the number of words in message *j* of student *k*, ML can be defined as

$$ML_i = \frac{\sum_j NW_{ij}}{\sum_k \sum_j NW_{kj}} \quad \cdots\cdots (2).$$

Message Count

Message Count (MC) measures how often a student participates in the class. It reflects the activeness of the student's class participation. MC of a student is defined as the proportion of messages that are generated by the student. MC_i is used to represent the message count of student *i*.

Performance Indicator

Taking together the three measures discussed above, we define the Performance Indicator (PI) as

$$PI_i = \alpha KD_i + \beta ML_i + \gamma MC_i \quad \cdots\cdots (3)$$

where PI_i is the performance indicator score assigned to student i, and the coefficients α, β, γ are the weights of the three measures respectively. The coefficients are adjustable. Instructors can define the values by specifying the importance of each of the three evaluation aspects. For example, if an instructor emphasizes on the quality of students' work most, and the quantity least, he might define the weights *as* α *=3,* β *=1, and* γ *=2.*

EXPERIMENT

To validate the model, we select 2 courses from the CIS department at NJIT, one in management domain (*C1*) and the other in information science domain (*C2*). Both courses are supported by WebBoard (http://webboard.njit.edu), an electronic system that allows instructors and students communicate with each other via text message exchanges asynchronously. In WebBoard, messages are organized in a tree structure. Participants can post a new topic in a conference, or reply to an existing message. The messages are stored in database in HTML format.

A program is written to download the course messages. It simulates the web browser by sending appropriate HTTP request to the WebBoard system, and finds information of interests from the returned HTTP response. The course messages are organized in its original structure, and other information, such as the post date and the author's name, is also saved. The HTML messages are then converted to plain text format, with unneeded information - message headers, signatures, and quoted lines, etc. - removed.

The 2 courses differ in their purposes of using the system. In *C1*, the instructor gives 8 discussion topics, and for each topic, students have to respond with an original post and reply to others before the deadline. The topics are selected to cover the major issues discussed in the class. In contrast, *C2* has only one discussion conference, and there are no specific discussion topics. Students are encouraged to find interesting resources to share with the class. It is used to evaluate students' online participation, which is counted as a small portion of the final grades. So, the actual grades compared to are the overall grades for C1 and the participation grades for C2.

To calculate the keyword density, we implement a noun phrase extractor that identifies noun phrases from free text (Wu and Chen, 2003). The counts of individual words and noun phrases for each message, each student, and the class are computed. Because we do not know the instructors' grading preferences, in this study we simply set all the coefficients α, β, γ to 1, assuming that they are equally important, though a more accurate setting would produce better results.

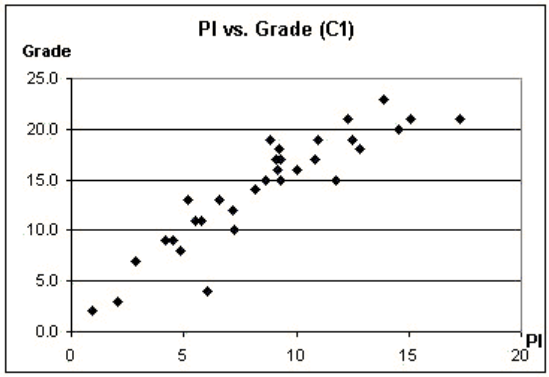
Copyright © 2004, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Table 1: Summary of the PI scores

	Range (%)	N*	Mean (Std.)
C1	0.92 ~ 17.3	32	8.6 (3.9)
C2	0 ~36.5	27	8.6 (11.3)

**: Number of students*

Figure 1: Correlation between the PI Scores and the Actual Grades for C1



RESULTS

The PI scores for the two courses are summarized in Table 1.

To examine how well this measure reflects the real performance of students, we calculate the Pearson product-moment correlation (*r*) between the PI scores and the actual grades assigned by the instructors. The correlation is 0.91 for *C1*, and 0.86 for *C2*. According to the reports from the literature, for various kinds of essay grading, the correlation between human judges varies from 0.5 to 0.9 approximately. It is reasonable to assume that the correlation between human judges in class grading also falls into this range. The results suggest that the computer grader performs well in terms of correlating with human evaluators, and the performance indicator reflects the students' actual performance at a high degree. Figure 1 shows the relationship between the PI scores and the actual grades for *C1*.

Unlike the essay grading approaches, which attempt to assign a concrete score to an object, our performance indicator score alone cannot predict the actual grade a student deserves. However, as mentioned above, because the PI scores highly correlate with the actual grades, they do distinguish "good" students from "poor" students. The rank order of the students by the PI scores would be more interesting to instructors. We compute another correlation. First, students are ranked by their actual grades in descending order, and the rank order is recorded as *Rg*. Similarly, another rank order, *Rpi*, is obtained by ranking the students by their PI scores. The Spearman rank-order correlation between *Rg* and *Rpi* is computed (0.92 for *C1* and 0.91 for *C2*). The results are shown in Table 2.

The high correlation between the rank orders shows that the PI scores rank students correctly according to their performance. For comparison, the two rank orders of *C1* are shown in Figure 2.

DISCUSSION AND CONCLUSIONS

Automated student performance evaluation works surprisingly well. For both the absolute scores and the relative rank orders, correlations between the PI scores and the actual grades are generally high. The result indicates that a model combining content and text feature analysis can be used for automated evaluation of students' performance.

Evidence of the supplementary role of the computer grader is also found from 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

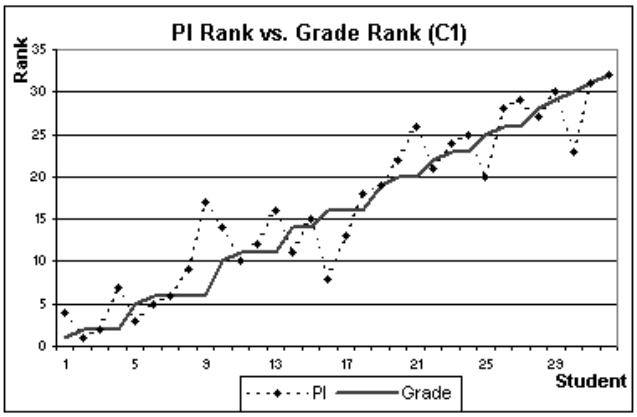
Table 2: Correlations

	r	r_s
C1	0.91	0.92
C2	0.86	0.91

r: *Pearson product-moment correlation between the PI scores and the actual grades*

r_s: *Spearman rank-order correlation between rank orders*

Figure 2: Rank Orders of the PI scores and the Actual Grades for C1



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 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 this to be less effort. Therefore, the original grade was confirmed. Having the program serving as a second grader will help the instructor to capture the outliers, and to reduce the misjudgment, bias, or errors in grading.

In this preliminary study, only two courses are selected for evaluation. This limitation prevents the model from being generalized to courses of other formats and in other domains. We plan to conduct the experiment with more different courses.

Another research possibility is that, even though we examined the correlation between the PI scores and the actual grades, the contribution of individual measures to the PI score is still unknown. The next step will be exploring the relationship at the individual measure level by asking instructors their grading preferences, e.g. a=5, b= 3, and g=2 might mean the instructor would like to give higher graders to students who are more able to synthesize knowledge learned from the class, rather than post many short content-poor messages.

REFERENCE

Bachman, F. L., Carr, N., Kamei, G., Kim, M., Pan, M. J., Salvador, C., Sawaki, Y., A reliable approach to automatic assessment of short answer free responses, Proceedings of COLING 2002, The 19th International Conference on Computational Linguistics.

Burstein, J., Kaplan, R., Wolff, S. and Chi 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.

Burstein, J., Kukich, K., Wolff, S., Chi Lu, Chodorow, M., Braden-Harder, L. and Mary Dee Harris, Automated Scoring Using a Hybrid Feature Identification Technique, in Proceedings of the Annual Meeting of the Association of Computational Linguistics, Montreal, Canada, 1998.

Copyright © 2004, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Burstein, J., Leacock, C. Chodorow, M., 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

Foltz, P. W., Laham, D., and Landauer, T. K., The Intelligent Essay Assessor: Applications to Educational Technology, Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2), 1999.

Hiltz, S. R., The Virtual Classroom: Learning Without Limits Via Computer Networks, Norwood NJ, Ablex, 1994.

Jones, E. L., Grading student programs - a software testing approach, The Journal of Computing in Small Colleges, Volume 16 Issue 2, January 2001

Kamp, H. A, Theory of Truth and Semantic Representation, Formal Methods in the Study of Language, Vol. 1, (J. Groenendijk, T. Janssen, and M. Stokhof Eds.), Mathema-tische Centrum, 1981.

Landauer, T., Laham, D., Rehder, B., and Schreiner, M., How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, 1997.

Landauer, T. K., Foltz, P. W., and Laham, D., An introduction to latent semantic analysis. Discourse Processes, 25, 259-284, 1998.

Landauer, T. K., and Psotka, J., Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA, Interactive Learning Environments, 8(2) pp. 73-86, 2000.

Larkey, L. S., Automatic essay grading using text categorization techniques. Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 90-95, 1998.

Page, E. B. The imminence of grading essays by computer. Phi Delta Kappan, 238-243, January 1966.

Page, E. B., Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62, 127-142, 1994.

Page, E. B. and Petersen, N. S., The computer moves into essay grading, Phi Delta Kappan, March, 561-565, 1995.

Snow, C. E. and Ferguson, C. A. (Eds.) Talking to Children: Language Input and Acquisition, Cambridge, Cambridge University Press, 1997.

Wu, Y- B., and Chen, X., Extracting Features from Web Search Returned Hits for Hierarchical Classification. Proceedings of International Conference on Information and Knowledge Engineering. (pp103-108). Las Vegas, NV, 2003.

Copyright © 2004, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

0 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/proceeding-paper/automated-evaluation-students-performance-analyzing/32299

Related Content

Two Rough Set-based Software Tools for Analyzing Non-Deterministic Data
Mao Wu, Michinori Nakataand Hiroshi Sakai (2014). *International Journal of Rough Sets and Data Analysis (pp. 32-47).*
www.irma-international.org/article/two-rough-set-based-software-tools-for-analyzing-non-deterministic-data/111311

ICT and Knowledge Deficiency
Rosa Iaquinta (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 4575-4582).*
www.irma-international.org/chapter/ict-and-knowledge-deficiency/112899

Estimating Overhead Performance of Supervised Machine Learning Algorithms for Intrusion Detection
Charity Yaa Mansa Baidoo, Winfred Yaokumahand Ebenezer Owusu (2023). *International Journal of Information Technologies and Systems Approach (pp. 1-19).*
www.irma-international.org/article/estimating-overhead-performance-of-supervised-machine-learning-algorithms-for-intrusion-detection/316889

Dynamic Interaction and Visualization Design of Database Information Based on Artificial Intelligence
Ying Fan (2023). *International Journal of Information Technologies and Systems Approach (pp. 1-13).*
www.irma-international.org/article/dynamic-interaction-and-visualization-design-of-database-information-based-on-artificial-intelligence/324749

Representations, Institutions, and IS Design: Towards a Meth-Odos
Gianluigi Viscusi (2012). *Phenomenology, Organizational Politics, and IT Design: The Social Study of Information Systems (pp. 131-141).*
www.irma-international.org/chapter/representations-institutions-design/64681