Assessing Student Learning Through Keyword Density Analysis of Online Class Messages

Xin Chen
New Jersey Institute of Technology
xc7@njit.edu

Brook Wu
New Jersey Institute of Technology
wu@njit.edu

ABSTRACT

This paper presents a novel approach that automatically assesses student class performance by analyzing keyword density in online class messages. Class performance of students is evaluated from three perspectives: the quality of their course work, the quantity of their efforts, and the activeness of their participation; three measures - keyword density, message length, and message count, are derived from class messages respectively. Noun phrases are extracted from the messages and weighted based on frequency and length. Keyword density is defined as the sum of weights of noun phrases appearing in a student's messages. Message length and message count are the number of words and the number of messages in the student's messages respectively. The three measures are combined into a linear model, in which each measure accounts for a certain proportion of a final score, called performance indicator. 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 the performance indicator score are highly correlated with those by the actual grades as well. Evidences of the supplemental role of the computer assigned grades are also found.

Keywords

Keyword Density, Learning Assessment, Distance Learning, Noun Phrase

INTRODUCTION

Advances in information technology have populated the online delivery of various types of classes, so called virtual classrooms (Hiltz, 1994), in which instructors and students interact with each other mainly by exchanging messages in natural language expression. Students' work submitted online accounts for a large portion of their final grades. To assess students correctly, instructors have to read through all their messages and other online submissions, which can easily accumulate to a large volume over a course of one semester. In addition, instructors may have different grading preferences, which will introduce bias in grades. Combining the assessments from multiple judges has been proved useful to increase the correctness of the final decision (Winkler and Clemen, 2002), so in addition to the instructor who solely makes the judgments of student performance, a second grader is useful and necessary. Using a second human grader, however, is not feasible in most situations because of the cost and limit of human resources. Computer programs, therefore, could be a better alternative. They could supplement to the judgments made by the instructor, and help instructors improve their grading by alerting them with disagreements to the grades they assign.

The idea of developing computer programs to grade student's work was introduced in 1960s (Page, 1966). It has been a long-time interest to researchers, and has been used widely in assessing various types of students' works, such as computer programs (Jones, 2001), prose (Page, 1994), language tests (Bachman et al, 2002), and essays (Page and Petersen, 1995; Larkey, 1998; Foltz et al, 1999; Landauer and Psotka, 2000; and Burstein et al, 2003). However, although the goal of these approaches is to automatically assess student learning, most efforts have been focused on individual work assessment (e.g. an essay, a piece of program, etc.), despite that students' course work accumulates over time. Correct assessment requires taking the entire volume of submissions into consideration. Recently the problem of text mining, a.k.a. text data mining (TDM) (Hearst, 1999) or knowledge discovery from text (KDT) (Feldman, 1995), is attracting increasing attention. With the large amount of class discussions and other documents (e.g. assignments, essays and projects) in text format, text mining techniques could be a solution for learning assessment.

This paper addresses the problem of automated assessment of student learning within a period of learning process, and presents a novel approach, which assesses student learning by mining keyword density in their class messages. Our approach is, first, to extract conceptual entities, called keywords, from the class messages. Next, keywords are weighted according to their lengths and occurrences in the messages. The weight of a keyword reveals the importance of the corresponding concept.

Message length and message count are two other measures, and refer to the number of words and the number of messages respectively. The three measures are combined into a linear model, in which each measure accounts for a certain proportion of a final score, called performance indicator.

The remainder of this paper is organized as follows. Section 2 presents the basic terminology, notation, and concepts concerning keyword density and the assessment model. In section 3, we discuss the experiment and evaluation method. Section 4 is the results and our analysis. Section 5 concludes the paper with some final remarks.

ASSESSMENT MODEL

The basic idea in this work is to access student learning by analyzing the messages and documents produced in the supporting e-learning systems. This section presents the assessment model, including the basic concepts concerning keyword density, 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; the proposed three measures - keyword density, message length, and message count, are derived from the class messages to measure each assessment aspect respectively.

Keyword Density Mining

Keyword density measures the quality of a student's work by analyzing the contents of the student's messages. Similar content-based approaches are found in essay grading literature. They focus on the semantic relationships between words and the context. Manually graded training essays are required to construct a semantic space, which reflects the contextual usage of words. A test essay is compared to the documents in the space, and assigned with a score according to the grades of the nearest essay(s). Examples are the early work in Educational Testing Services (ETS) (Burstein et al, 1996) and the Latent Semantic Analysis (LSA) model (Landauer et al, 2000). However, the major characteristic of our approach is that it does not require training essays, which is expensive and nearly unavailable in our problem. In addition, our approach deals with not just one essay, but a set of class messages. Below we give the basic definitions of the concepts used throughout of the paper.

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.

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. Therefore, noun phrases are considered the conceptual entities in text messages. We define keyword as a simple, non-recursive noun phrase, i.e. basic noun phrase.

Identifying basic noun phrases from free text is well researched in the field of Natural Language Processing. We implement a noun phrase extractor, which combines both Markov-based (Church, 1988) and rule-based (Brill, 1992) approaches. It disambiguates a multi-part-of-speech (POS) 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 (Fellbaum, 1998), which contains words divided into four categories (noun, verb, adjective, and adverb) and the number of senses of each word used in one of the categories, is used to assign the right POS tag. The initial POS tag for a word 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 every 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 to disambiguate the multi-tag words. 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 extracts noun phrases by selecting the sequence of POS tags that are of interest. The current sequence pattern is defined as [A] {N}, where A refers to Adjective, N refers to Noun, [ ] means optional, and { } means repetition.

Keyword Weighting Scheme

In previous studies (Chen and Wu, 2004; Wu and Chen, 2004), keywords (noun phrases) are assumed equally important. In this study we borrow the idea of term weighting in information retrieval (Salton and Buckley, 1988) to assign different weights to keywords. The importance of a keyword is measured by its frequency. The more frequently a keyword is used, the more important it is. However, if a keyword is used by more students, it becomes less important in terms of differentiating one student's contribution to the class concept base from others'. In other words, a student fully contributes to the class concept base by only 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 ones. For instance, in a course of Information Systems, *information systems* is likely to be used by most of the students, while *expert systems* is 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 learnt in the class 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 brining in new concepts, not just repeating the existing ones.

The length of a noun phrase should also be taken into consideration. Longer noun phrases are more descriptive than shorter ones. As we can see from a real example from 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 assign weights to keywords as follows.

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

where 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 measure in Information Retrieval (Salton, 1989), but the inverse frequency of a keyword is across students, not messages.

Keyword Density

After assigning weights to keywords, we calculate Keyword Density (KD) by adding up the weights of keywords contributed by each student. The result is divided by the sum of weights of all keywords in the concept base. It is denoted as follows.

$$KD_i = \frac{W_i}{W},$$

where KD_i is the keyword density for student i, W_i is the sum of weights of the keywords contributed by student i, W is the sum of weights of all keywords.

Message Length

The quantity of a student's efforts is measured by the length of his/her messages (ML). ML is calculated by counting all the words (not noun phrases), no matter duplicated or not, in the student's messages. The absolute number is proportioned by the total 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, we have

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

Message Count

Activeness of participation could be measured by the log on times, but this information is not available inside the text messages. If we consider posting a message a valid participation, participation frequency can be measured by the Message Count (MC), which is the number of messages posted by a student. MC is defined as

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

where MC_i is the message count of student i, and n_i is the number of messages posted by student i.

Assessment Model

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

$$PI_i = \alpha KD_i + \beta ML_i + \gamma MC_i,$$

where PI_i is the performance indicator score assigned to student i, and the coefficients α, β, and γ 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 α=5, β= 3, and γ=1, the instructor would like to give higher grades to students who are more able to synthesize knowledge learned from the class, rather than to post many short content-poor messages.

EXPERIMENT

Five classes are selected for model validation. All classes are supported by an electronic conference system that enables instructors and students to communicate with each other via text message exchanges asynchronously. The class information and their conference design are summarized in table 1.

ID	Domain	Conference Design
C1	Management	Required discussions on major topics in the course. Students are graded according to their posts only.
C2	Information Science	Optional discussions on topics that students would like to share with each other. Grades are assigned to students' online participation.
C3	Information Systems	Activities include required class discussion, debates, oral presentation, assignment submission, course project, and optional discussions. Some of the grading items are judged according to the documents submitted online, while others are based on materials handed-in such as project reports, PowerPoint slides, and so on.
C4	Information Systems	
C5	Information Systems	

Table 1: Summary of Course Information

All class messages are downloaded and converted to plain text, from which keywords are extracted and weighted. When calculating the performance indicator score, we set the three coefficients, α, β, and γ, to 1, because we do not know the instructors' grading preferences. However, when the grading preferences are known, it is easy to adjust the coefficients to comply with the grading preferences.

RESULTS AND ANALYSIS

Table 2 shows some selected keywords with the weights and frequency. Although *business* has a high frequency (30), its weight is still 0 because it is used by all students. The results also show that longer phrase and more frequent phrases have higher weights.

Keyword	Weight	Frequency/# of Authors
business	0.00	30/13
business activities	5.15	1/1
business application technology	6.39	1/1
business models	9.88	3/3

Table 2: Selected Keywords and Their Weights

Table 3 summarizes the PI scores of the five classes.

	Range	N*	Mean (Std. Dev.)
C1	0.01~0.23	32	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)

*: Number of students

Table 3: Summary of the PI scores

To examine how accurately the PI scores assess student learning, we calculate the Pearson product-moment correlation (r) between the PI scores and the grades assigned by the instructor. Correlations between the individual measures and the actual grades are also calculated (see table 4 for details). The results show that there is a high correlation between the PI scores and the actual grades (from 0.57 to 0.92). According to the reports in the essay grading literature (Page and Petersen, 1995; Larkey, 1998; Foltz et al, 1999; Landauer and Psotka, 2000; and Burstein et al, 2003), correlation between the judgments of human expert varies from 0.5 to 0.9 approximately. It is reasonable to assume that correlation between grades of two instructors also falls into this range. The results, therefore, suggest that the computer grader performs well in terms of correlating with human evaluators.

	$r_{PI\text{-}G}$	$r_{KD\text{-}G}$	$r_{ML\text{-}G}$	$r_{MC\text{-}G}$	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

$r_{KD\text{-}G}$: Correlation between the KD 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

Table 4: Correlations

The low correlation in C3 and C5 catches our attention. By looking into the conference boards, we find the following possible reasons:

- The two courses contain a few conferences that are designed for course administration. For example, C3 has three conferences for assignment distribution and submission. The questions and answers regarding the assignments should have been excluded from analysis. Another similar conference is the exam conference. If the content of a conference is not used by the instructor for grading but processed by our program, noises will be introduced.
- 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/closed conferences that cannot be processed by our program.

Unlike essay grading approaches, which attempt to assign a concrete score to an object, our performance indicator score does not attempt to predict the actual grades of students. It assesses students by comparing them with each other to distinguish "good" students from "poor" ones. The rank order of students by the PI scores would be more interesting to instructors. It could help instructors assess students by grouping them at different levels, which is still a common approach used in practice

by many instructors. 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 results are shown in the last column of Table 4. The high correlation between *Rg* and *Rpi* suggests that the PI scores rank students correctly.

Evidence of the supplementary role of the PI scores 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 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 this to be less effort. Therefore, the original grade was 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. When consulted, the instructor explained that S is an exception in that class. S submitted almost every assignment late because of family and personal medical problems. The instructor accepted S' late assignments but gave low grades. After the makeup exam, the instructor changed S' final grade to a higher one. Because of 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.

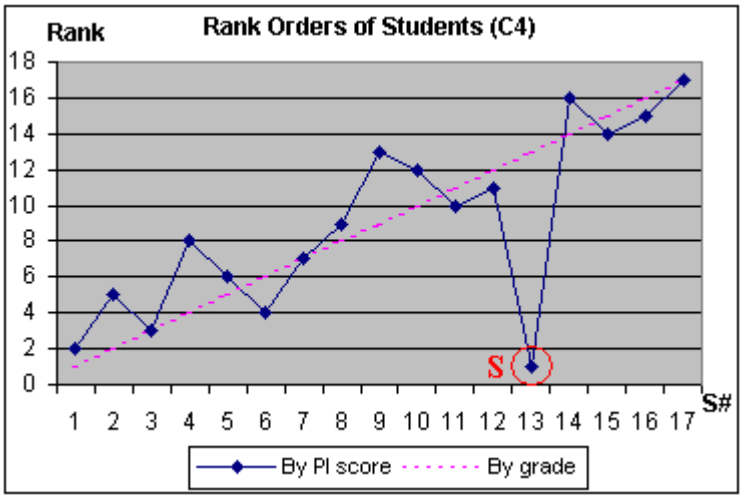

Figure 1: Rank Orders of Students by PI Score and by Grade (C4)

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

CONCLUSION

We have presented a model for automated assessment of student learning in virtual classrooms. This model is based on a text mining technique - keyword density mining from large body of texts. Noun phrases are extracted from class messages and weighted to build a concept base. Keyword density is then calculated for each student, and is further combined with two other measures, message length and message count, into a linear model which predicts a performance indicator score for each student. The experiment shows that the assessment model works well. For both the absolute scores and the relative rank orders, correlations between the PI scores and the actual grades are generally high.

Our future research plan includes enhancing the current program, so that it can handle attachment files, and deal with special conferences. We also plan to test the model against courses in other domains.

REFERENCES

1. 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.
2. Brill, E.: A simple rule-based part of speech tagger. Proceedings of the Third Annual Conference on Applied Natural Language Processing, ACL. 1992.
3. Burstein, J., Kaplan, R., Wolff, S. and Lu, C., 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.
4. 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
5. Chen, X., Wu, B., Automated Evaluation of Students' Performance by Analyzing Online Messages, IRMA 2004 New Orleans
6. Church, K. W., A Stochastic Parts Programs and Noun Phrase Parser for Unrestricted Text. In: "Proceedings of ANLP-88", Austin, TX, USA, 1988.
7. Feldman, R. and Math, I. D. Knowledge Discovery in Textual Databases (KDT), In Proceedings of the First International Conference on Knowledge Discovery, KDD-95
8. Fellbaum, C. D. WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998
9. 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.
10. Hearst, M. A. Untangling Text Data Mining, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999
11. Hiltz, S. R., The Virtual Classroom: Learning Without Limits Via Computer Networks, Norwood NJ, Ablex, 1994.
12. Jones, E. L., Grading student programs - a software testing approach, The Journal of Computing in Small Colleges, Volume 16 Issue 2, January 2001
13. 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.
14. 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.
15. 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.
16. Page, E. B. The imminence of grading essays by computer. Phi Delta Kappan, 238-243, January 1966.
17. Page, E. B., Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62, 127-142, 1994.
18. Page, E. B. and Petersen, N. S., The computer moves into essay grading, Phi Delta Kappan, March, 561-565, 1995.
19. Salton, G. and McGill, M. J. Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1983
20. Salton, G. and Buckley, C. Term weighting approaches in automatic text retrieval, Information Processing and Management, vol. 24, no. 5, pp. 513--523, 1988.
21. Snow, C. E. and Ferguson, C. A. (Eds.) Talking to Children: Language Input and Acquisition, Cambridge, Cambridge University Press, 1997.
22. Winkler, R. L. and Clemen, R. T., Multiple Experts vs Multiple Methods: Combining Correlation Assessments, Decision Analysis, November 2002
23. Wu, Y. B. and Chen, X., Assessing Distance Learning Student's Performance: A Natural Language Processing Approach to Analyzing Online Class Discussion Messages, ITCC 2004, Las Vegas