T. Hirashima et al. (Eds.) (2011). Proceedings of the 19th International Conference on Computers in Education. Chiang Mai, Thailand: Asia-Pacific Society for Computers in Education

Analyzing Students' Usage of E-Learning Systems in the Cloud for Course Management

Tuan-Anh DOAN[1*] Jie ZHANG[1] William Chandra TJHI[2] Bu Sung LEE[1]
[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]A*STAR, Institute of High Performance Computing, Singapore
*tadoan@ntu.edu.sg

Abstract: E-learning systems are now considered a core IT support in most Institute of Higher Learning. Each student is assigned a set of courses based on their preferences for each semester. Learning materials are posted and evaluated through the system. Mining access and usage log data of the e-learning system can give insights such as on how these materials are accessed by students in their order of preferences on each of the courses. A study was carried out, and the results show that there is good correlation between students' performance and e-learning usage. In this paper, we describe our methodology on analyzing students' usage of e-learning systems for course management with the help of cloud computing services. We present a student performance predictive model built from the retrieved logs that is confirmed to achieve sufficiently good accuracy.

Keywords: Course Management, E-Learning, Data Analytics, Cloud Computing

Introduction

Modern technologies have transformed students learning methods. Instead of gaining most of knowledge from classes, they can now acquire knowledge from many resources scattered across the Internet. Many e-learning systems have also been implemented for online education purposes [6], e.g. Blackboard deployed by Nanyang Technological University. A crucial element for these systems is effective course management. Course managers need to make informed decisions about what materials should be the most appropriate to be presented to students (learners) and what learning strategies or methods should be used for the students [2]. Most existing commercial course management systems provide only basic information about student access statistics to the e-learning system. To improve this, students' learning behaviors need to be modeled by taking into account different factors from distributed information sources and multiple domains. In our previous work [7], a framework for user-driven data analytics in the cloud has been proposed to support an effective course management system. This framework provides a roadmap towards provision of various data analytics services to course managers, including an intelligent crawler to find relevant data, a meta miner to recommend the best workflow together with the transfer learning for producing different models (such as student models and course material/learning object models) possibly across different domains, the cloud compute service to support computation and storage in heterogeneous and distributed environment, and the visual analytics service to allow course managers for interaction with the models built. In particular, a usage miner is provided by the framework to mine the usage patterns of students on applications (i.e. e-learning systems) of the framework.

In this paper, we provide a case study of the cloud analytics framework. We describe our method in processing students' log data in the cloud environment to extract students' sessions and analyze those sessions to predict students' performance. Students' log data collected each semester can easily reach hundreds of gigabytes. We harness the cloud computing power to process these data in our experiments. The Hadoop MapReduce framework in particular is used to extract students' sessions from 171GB of log data. Each session represents the list of activities when students access the e-learning system from the

time the students log in until the time they log out or become inactive. We then model the student's learning manner by a set of attributes. Those attributes are used as inputs to Weka, a platform containing a set of algorithms for data mining tasks. Five classification methods are selected from Weka to predict students' performance. Two main results and contributions have been drawn from our work. Firstly, the cloud computing significantly speed up the process of analyzing students' log data. Secondly, the prediction results using the attributes we extract from students' log data provides sufficiently good accuracy. Depending on the predicted students' performance, course managers can then make decisions about which students need help from extra materials and other students.

1. Analyzing Students' Usage of E-Learning System

As the first and most important step towards our course management system, we analyze students' usage of e-learning systems and predict their performance for the courses, consisting of four phases: data collection, data preprocessing, modeling and evaluation.

1.1 Data Collection

The data used in this study are real log data collected from the Nanyang Technological University e-learning system called "Edventure" from August 3 to December 23, 2010 (semester 1, academic year 2010-2011). The e-learning system is used by more than 30,000 students comprising multiple courses. The full dataset uncompressed is about 171GB. We select three courses from School of Computer Engineering for predicting students' performance. The first course has 24 students, the second has 126 students and the third has 23 students. We obtain the students' performance in these three courses for our study. Because of privacy issues, instead of dealing with the real students' grades, we categorize students into four groups and predict the performance solely base on those groups. The methods we use to categorize students will be described in details in the modeling phase.

1.2 Data Preprocessing

In this phase, all the records in log data that have the same IP address and session ID were grouped into one session. Thus, we assumed that different sessions have either different IP addresses or different session IDs. Afterwards, the whole dataset was uploaded into Hadoop Distributed File System (HDFS). We then wrote a Java application in Hadoop MapReduce framework to extract all the students' sessions. It took nine hours and fifteen minutes to process 171GB of log data in our Hadoop cluster. The Hadoop cluster consists of 4 data nodes. Each node deploys on a CPU Intel Xeon 2.4 GHz with 1GB of RAM while our PC uses Intel Core 2 at 2.67GHz and has 2GB of RAM. The running time is non-distinguishable on 50MB of log data (32 seconds on Hadoop cluster and 33 seconds on PC). However, as the data size increases, Hadoop cluster's running time becomes much shorter than PC. At 1GB of data, our PC takes 8580 seconds (more than two hours) to process while a Hadoop program only takes 152 seconds (less than three minutes) to finish running. Cloud computing is proven to be able to significantly speed up the data processing. After obtaining all the students' sessions in the e-learning system, we extract from there only those sessions which include activities on the three courses we discussed in the data collection phase.

1.3 Modeling

The purpose of this phase is to model the students' learning behavior in a way that could yield good precision when we apply classification algorithms on evaluation phase. To protect the students' privacy, we use four categories to classify students' performance:

- EXCELLENT: if grade is A+, A or A-;
- GOOD: if grade is B+, B or B-;
- AVERAGE: if grade is C+ or C;
- BELOW AVERAGE: if grade is D+, D or F

Moreover, since we are interested in determining which students have problems in understanding the course and thus may need extra learning materials or help from other good students, we also propose two other ways to categorize students' performance. The first way has three categories which merges AVERAGE and BELOW AVERAGE students into one category (called AVERAGE). The second way only has two categories. Besides, merging AVERAGE and BELOW AVERAGE students as in the first way, it also groups EXCELLENT and GOOD students into one category (called ABOVE AVERAGE). Afterwards, based on the students' usage of the e-learning system, we devise five attributes extracted from the logs that attempt to model the students' learning behavior. Table 1 compiles all attributes used as features for classification algorithms.

Table 1. Attributes used by each student

Name	Description
totalTime	Total time spent in e-learning system
freq	Number of times accessed
numDocs	Number of materials accessed
numDiscussions	Number of times accessed discussion board
numElearning	Number of times viewed e-learning videos

The number of hours which the students attend class has been shown to be highly correlated with academic success [5]. Thus, we consider the total time spent on the e-learning system as an implication of student behavior. Furthermore, students who believe that the course requires regular homework are those who more likely to succeed [5]. Therefore, we use the frequency of access as an indicator whether students spend time regularly to study. The number of materials students accessed during semester, and the number of times they accessed discussion board and viewed e-learning videos are the attributes that are aimed at modeling the active level of students on e-learning systems.

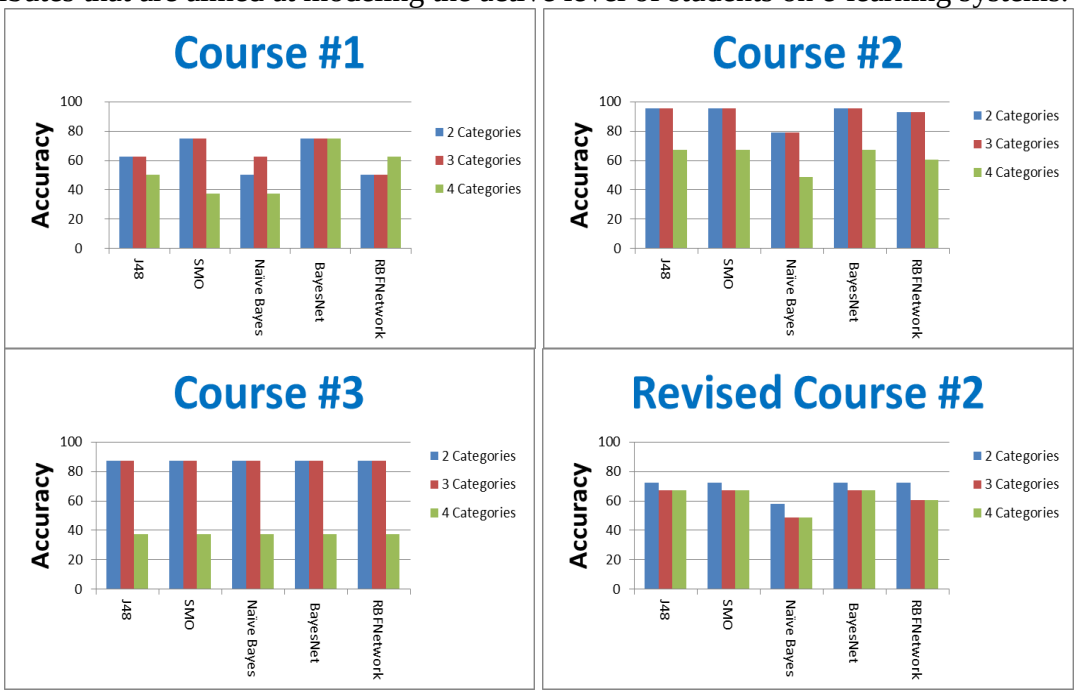

Figure 1. Accuracy of classifiers of three courses and revised course #2

1.4 Evaluation

Those attributes obtained in modeling phase are used as inputs to Weka. Five classification algorithms are selected to predict students' performance: J48, SMO, NaïveBayes, BayesNet and RBFNetwork. Figure 1 gives us the percentage of correctly classified students'

categories of three courses which have 24, 126 and 23 students accordingly. When we separate students into 4 categories, the accuracy of classifiers varies from 37.5% to 75.0%. On the other hand, the total rate of correct classification for 2 and 3 categories both range from 50.0% to 95.3%. In general, course #2 which has the highest number of students (126 students) gives higher accuracy when comparing to course #3 (23 students). A specific treatment of course #2 result was needed as most of the students in this class belong to either EXCELLENT or GOOD category. Hence, we decide to group AVERAGE and BELOW AVERAGE students into one category in three categories case and group GOOD, AVERAGE and BELOW AVERAGE together in two categories case. Figure 1 also shows the results of course #2 after modifications. The adjustment results in lower but more realistic percentage of correctly classified instances of two and three categories.

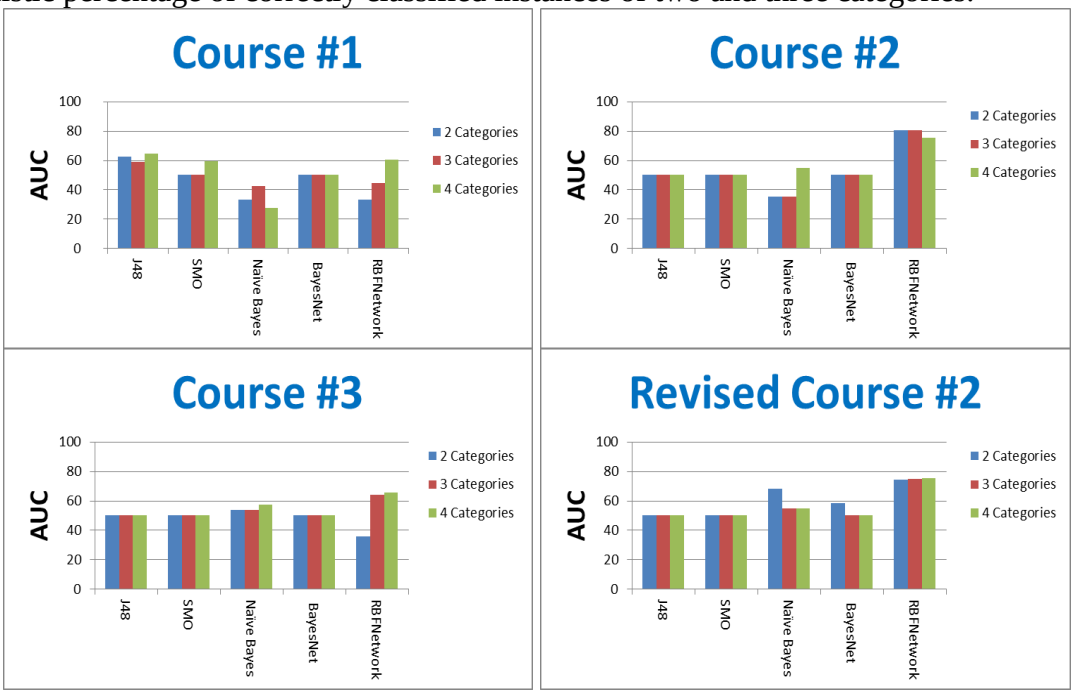

Figure 2. Area under the ROC curve (AUC) of the three courses and revised course #2

Furthermore, we also take into account the problem of imbalanced data. For instance, the distribution of students in course #2 is: EXCELLENT 23.01%, GOOD 73.80%, AVERAGE 3.17%, BELOW AVERAGE 0%. In order to reflect more accurately the performance evaluation of classification on imbalanced data, we use the Area under ROC Curve (AUC) to complement classification accuracy in Figure 1. The measure of accuracy becomes biased when classes are imbalanced. A clear illustration of the bias is in the case of the 2-categories classification of the course #2, in which the composition of the two classes is 96.81%-3.17%. In this case, a classifier that always considers any sample as EXCELLENT-GOOD would score 96.81% accuracy. On the other hand, the ingredients of AUC are the true positive and false negative rates. These rates are normalized on a per-class basis (i.e. they are "recall" instead of "precision" measure). Hence, any class-size skew is moderated by this normalization. To contrast with the accuracy measure, the AUC of a classifier that always assigns any sample to one of the classes is 0, as the ROC curve of this classifier is a single dot/point at either of the two extreme ends of the random diagonal ROC curve [4]. Figure 2 shows the AUC of the three courses and course #2 after modification. The AUC on multiclass case is calculated as the arithmetic mean of all the AUCs of available classes. The new result presents more conservative evaluation compared to the accuracy scores in Figure 1. While the AUC reveals the challenge in discriminating imbalanced classes by the classifier, some classifiers such as J48 in course #1, RBFNetwork in course #2 and revised course #2 or NaiveBayes in Course #3 still give a good AUC

ranging from 0.5 to 0.8. Moreover, feature analysis was performed to discover which attributes are more important than the others in classifying students' performance across three courses. We use attribute selection in Weka and choose ranker, information gain or chi-square and record down how many times an attribute appears in top of the rank list. Attributes *totalTime* and *freq* appear most of the time which suggests that those students who spend more time on the e-learning system and check the course contents regularly are more likely to succeed.

2. Related Work

There is a rich history of research on educational data mining. Interested readers can refer to Romero's survey [3] for a more complete reference. Lately, there have been several efforts on exploring factors that can predict academic performance. In [1], personality traits such as neuroticism, extraversion, openness to experience and academic behaviors such as absenteeism, essay-writing, and seminar behavior have large effects on academic performance. In [5], authors propose some attributes that can help to determine the achievement of university students using data mining methods. The authors, however, collect students' data by distributing a questionnaire to their students. Such method depends largely on the honesty of the students. A more recent work [4] proposes an experiment to classify students based on the data they collect from their e-learning system. However, this work did not provide any mean to deal with vast amount of students' log data. In contrast, our study uses cloud computing service to support computation and storage. Furthermore, some of the attributes we use to model the students' behavior are novel.

3. Conclusion

This paper describes the context of supporting an effective course management system for better e-learning, based on a framework for user-driven data analytics in the cloud. It is a new scalable technology that complements users of analytics by retrieval, integration and summarization/visualization of relevant heterogeneous information from external sources and facilitates user interpretation, interaction and collaboration. Within this context, we have then presented students' performance predictive model based on the data collected from our e-learning system. The teacher could use our classification in order to analyze students' achievement and provide assistance when students need.

References

[1] Chamorro-Premuzic T., Furnham A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personalit*y.

[2] Champaign, J. & Cohen, R. (2010). A Model for Content Sequencing in Intelligent Tutoring Systems Based on the Ecological Approach and Its Validation Through Simulated Students. *Proceedings of the Ninth Florida Artificial Intelligence Research Symposium (FLAIRS)*. Daytona Beach, Florida.

[3] Romero C. & Ventura S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications* no.33, 135 - 146

[4] Romero C. & Ventura S. & Espejo P. G. & Hervas C. (2008). Data mining algorithms to classify students. *Proceedings of the 1st International Conference on Educational Data Mining (EDM)* 8-17

[5] Superby J.F, Vandamme J.-P & Meskens N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. *Workshop on Educational Data*.

[6] Vassileva, J. (2009). Towards Social Learning Environments, *IEEE Transactions on Learning Technologies, 1*(4), 199-214.

[7] Zhang J., Tjhi W. C., Lee B. S., Lee K. K., Vassileva J. & Looi C. K. (2010). A Framework of User-Driven Data Analytics in the Cloud for Course Management. *Proceedings of the 18th International Conference on Computers in Education (ICCE).*