Open Academic Analytics Initiative (OAAI)

Table of Contents

Summary of final Findings from Course Pilots during OAAI

During the spring 2012 & Fall 2012 semesters, the Open Academic Analytics Initiative (OAAI), led by Marist College, successfully deployed an open-source Learning Analytics solution, developed by the project during the fall, at two community colleges (Cerritos College and College of the Redwoods) and one Historically Black College and University (HBCU) (Savannah State University) as a means to further research in this emerging field. As of its ending date of January 21, 2013, the Open Academic Analytics Initiative (OAAI) has successfully achieved all of its major project outcomes including:

1. developed and deployed an open-source academic early alert system
2. released, under an open-license, our OAAI Predictive Model
3. published research on "portability factors" of predictive models from one academic context to another
4. published research on the impact of different intervention strategies on student performance
5. disseminated outcomes and research findings to supported adoption of analytics at scale

In addition to these achievements, OAAI was selected as a *"2013 Computerworld Honors Laureate"* in the category of Emerging Technology as well as a recipient of the *"2013 Campus Technology Innovators Award"* of which only nine were selected from 235 applications this year.

Project Findings and Impact

Our research findings related to both the "portability" of predictive models and the impact of our intervention strategies on student performance (i.e. course completion, content mastery and semester-to-semester persistence). Our research efforts were associated with three of our five major project outcomes (specifically Outcomes #2, #3, and #4) and were among our most significant accomplishments. We have published two papers in peer-review journals to date related to our initial findings and a third paper, to the forthcoming Journal of Learning Analytics, that will discuss our comprehensive research methodologies.

The following summarizes our key findings which are discussed in more detail in the following sections:

1. Predictive models can be "ported" from one academic context to another and retain most of their predictive power. The OAAI Predictive Model remained in the 60-85% accuracy range (depending on institution) when deployed in our course pilots.
2. Overall, our interventions had a positive and statistically significant ($p = .013$) impact on average course grades. This positive trend was also found among "low income" students although just under the statistically significance level ($p = .054$).
3. Overall, our interventions had a positive and statistically significant ($p = .001$) impact on content mastery. This positive trend, which was also statistically significant ($p = .023$), was also found when we considered only "low income" students.
4. We found that course completion rates were higher among our control group (91%) then our treatment groups (85%) which were also found to be a statistically significant trend. In analyzing this finding in more detail, we determined that students in our treatment groups may be withdrawing from courses earlier than those in our controls which may explain this apparent reversal in the expected outcome.
5. We also found that there was no statistically significant difference in semester-to-semester persistence rates between our controls and treatment groups. We believe this may be the result of the limited scope of our pilots and the fact that we were impacting, at most, on only one of several courses that students were taking during the spring semester.

These findings, particularly those related to the impact of our intervention strategies, directly relate to NGLC's overall goal of dramatically improving student outcomes in college completion and demonstrate the positive impact OAAI has had on this goal. Please contact our Josh Baron at josh.baron@marist.edu.

Project Overview

The Open Academic Analytics Initiative (OAAI) will develop, deploy and release an open-source ecosystem for academic analytics designed to increase student content mastery, semester-to-semester persistence and degree completion in postsecondary education. As a result, we expect to see increases in adoption of academic analytics, particularly among institutions using the open-source Sakai Collaboration and Learning Environment, in both the short- and long-term.

Academic or learner analytics has received significant attention within higher education, including being highlighted in the recently released 2011 Horizon Report (Johnson, Smith, Willis, Levine, & Haywood, 2011). This interest can, in part, be traced to the work at Purdue University which has moved the field of academic analytics from the domain of research to practical application through the implementation of *Course Signals*. Results from initial *Course Signal* pilots between fall 2007 and fall 2009 have demonstrated significant potential for improving academic achievement. Despite this early success, academic analytics remains an immature field that has yet to be implemented broadly across a range of institutional types, student populations and learning technologies (Baepler & Murdoch, 2010). It is also clear that analytics alone do little to help students succeed academically thus improving our understanding of best practices related to student interventions remains a critical issue (Arnold, 2010).

To further advance the field of academic analytics, the OAAI, through technical development efforts, analytical research, institutional pilots and exploratory studies will:

a. Release, under an open-license, a Sakai "Student Effort Data" (SED) API that will capture user activity data and expose it in a secure fashion for use by early alert analytics-based tools.

b. Research the "portability" of predictive models used in academic analytics to better understand how models developed for one academic context can be effectively deployed in another.
c. Release an "open-source" OAAI Predictive Model for academic success that can be deployed by other institutions and overtime, be enhanced through open-source community collaboration.
d. Advance our understanding of technology-mediated intervention strategies by investigating the impact that engagement in an "Online Academic Support Environment" has on student success.
e. Document best practices related to deploying academic analytics using Sakai and the Pentho open-source business intelligence suite, including models for collecting appropriate learner data from institutional student information systems.

Although focused on open-source solutions, the OAAI will also facilitate, through the release of the Sakai SED API and OAAI Predictive Model, broader use of proprietary early academic alert systems, such as IBM SPSS Decision Management for Student Performance and SunGard Higher Education Course Signals. This will further magnify our impact on adoption and ultimately student academic success.

More details can be found in the attached full *OAAI Grant Proposal.pdf* to the Next Generation Learning Challenges program.

How to Get Involved

Marist College is currently serving as the lead institution for the OAAI. If you are interested in getting involved, are looking for the latest project updates or just have general questions about academic analytics and Sakai, please contact Josh Baron, OAAI Principal Investigator, at Josh.Baron@marist.edu.

Published Papers, Articles, and Reports

- ***Mining academic data to improve college student retention: An open source perspective.pdf : A short technical paper which has published predictive modelling methodologies for OAAI*** – Dr. Eitel Lauria, Mr. Josh Baron and three Marist computer science graduate students (Ms. Mallika Devireddy, Ms. Venniraiselvi Sundararaju and Mr. Sandeep Jayaprakash) have had their paper, that discusses initial findings from comparing correlations in Marist's datasets with those found by Dr. John Campbell in his work at Purdue, accepted at the 2nd International Conference on Learning Analytics and Knowledge (LAK 2012).

- ***"OAAI: Initial Research Findings.pdf" –*** *A short technical paper which has published findings from Initial "Portability" Research at Partner instituions* – Dr. Eitel Lauria, Mr. Sandeep Jayaprakash, Dr.Erik Moody, Mr. Josh Baron and Marist computer science graduate students (Ms. Nagamani Jonnalgada) have had their paper, that discusses initial findings from comparing predictive model performances in Marist's datasets with the model performance when deployed at partner institutions of different academic context than Marist, accepted at the 3rd International Conference on Learning Analytics and Knowledge (LAK 2013).

- **"Open Academic Analytics Initiative: Final Progress Report.pdf"** submitted to NGLC grant committee at EDUCAUSE, details the achievements of the OAAI grant, crucial project findings and their impact in the higher education realm. OAAI Final Progress Report.pdf

- ***"Harnessing the Power of Technology, Openness, and Collaboration"*** by Josh Baron and Kimberlee Thanos in EDUCAUSE Quarterly Volume 34, Number 4, 12/15/2011.

- ***"Opening Up Learning Analytics for the Community College" : A Q&A with Josh Baron and JoAnna Schilling*** by Mary Grush appeared on the Campus Technology web site in December 2012.

OAAI Technical Documentation

Required Software Tools & Installation guides

- ***Microsoft SQL Server 2005/2008***
- ***Pentaho Data Integration (Kettle)***
  Download the latest version of Kettle from http://kettle.pentaho.com/
- ***Weka Data mining Tool***
  Please refer to the following installation guide.**WEKA Installation Files.zip**

- ***Weka Scoring plugin for Kettle***
  Refer **Steps to Add Weka Scoring Plugin to Kettle .docx**
- **Pentaho Report Designer**
  http://reporting.pentaho.com/report_designer.php

Open Academic Analytics Initiative Predictive Model

The primary focus of the OAAI predictive modelling is in

- Devise, develop and successfully deploy an open source predictive model based on a Data mining approach.
- Research into portability of the models for student performances in Learner analytics to understand how models developed for one academic context (eg. large research institution) can be deployed effectively into another(eg. community college).

Building a Predictive model using Data-Mining approach

Extraction Phase

The predictive model requires 4 different files providing adequate information about the student background, activity and his performance.

- Student Demographics Data (Personal)
- Course Data
- CMS Gradebook Data (Sakai Gradebook)
- CMS Usage/Events Data (Sakai Events)

The first 2 files are extracted from the Banner using a tool called ARGOS ,whereas the Gradebook and Events files are extracted from the CMS (in our case Sakai).
Please refer to the following document which gives detailed description regarding the Required Datasets Format ***Required Dataset format.docx***
For the queries required as a part of extraction process of Sakai Data (Gradebook and Events) refer the following document. ***Queries for Extraction of Sakai(CMS) Gradebook & Events data .doc***

Transformation Phase using Kettle

Once the 4 different datasets Demographics, Course, Gradebook, Sakai Events are obtained the next step would be clean the datasets and prepare them for the training process in predictive modelling. The diverse set of data gives rise to many data quality issues arising mainly due to

- Variability in Sakai tools usage (tools not used, data not entered, missing data)
- Variability in instructor's assessment criteria
- Variability in workload criteria
- Variability in period used for prediction (early detection)
- Variability in Multiple instance data (partial grades with variable contribution, and heterogeneous composition)

A predictive model is usually as good as its training data. Since there is a diverse set of data available it becomes essential to find appropriate relations among data and identify metrics and combine them in an effective way so that it can maximize the predictive power of the model. During the process of transformation of the datasets we try to improve the Data Quality by dealing with all the above mentioned issues with the creation of predictive metrics which helps us derive a uniform standard/ criteria among the variability.

The important technique employed to create this uniform standard is effective calculation and use of ratios for Gradebook and Sakai Events metrics mentioned earlier. For the entire enrollment in a course we calculate an Average percentage of Sakai usuage / Gradebook scores and then we calculate the ratio of Students Usage/Scores against the average course usage / Scores.

Hence the formula for Usage / Gradebook metrics are calculated using

- **Percent of usage over Avg percent of usage per course**
- **Effective Weighted Score / Avg Effective Weighted Score**

After cleaning each of the datasets separately and identifying and creating the required predictive metrics we merge all the datasets into a large single dataset through a series of Join operations and feed the training datasets it to the model and create a library of models.

Training datasets - Marist Undergraduate Fall 2010 , Marist Undergraduate Spring 2011.

Refer **Sample flows(Undergrad Fall 2010)** and documentation regarding the Transformation process using the following links ***Undergrad 10F kettle flows.zip*** ***ETL of datasets with Kettle flows.docx*.**

Sample flows(Undergrad Fall 2010) Video - ***OAAI Data Integration.wmv***

We follow the same steps to perform transformation and pre-process the data on Undergraduate Spring 2011.

Load Phase - Feeding cleaned datasets into WEKA machine learning tool for Training puposes.

Here we use full semester datasets for Gradebook and Events where we already know the students at RISK. In order to train the model we reverse engineer the datasets and try to identify as to how the students are categorized under academic risk. We try to find intricate details as to what goes into identifying Academic_Risk students using the predictive metrics we built in the transformation phase and will try to find the correlations between each metric using machine learning algorithms from WEKA. The knowledge flow developed in Weka was replicated in IBM SPSS modeler and identical model were trained. The predictive model trained in Weka was then verified and validated against the model trained in IBM SPP modeler for consistencies.

Refer the following document where the WEKA predictive model is explained.*Academic Analytics model- Weka flow.docx*

Weka analytics model flows - *Weka Analytics model flows.zip*

Sample Demo Video of the Weka Predictive model - *OAAI Predictive Modeling.wmv*

Exporting OAAI Preddictive model in PMML format

One of our main objectives of the Open Academic Analytics Initiative is to release the verified and validated Predictive analytics model under the standard PMML format. The Data-mining and Machine learning tool WEKA, while it supports import functionality of the PMML based Predictive models, it does not support export functionality. Hence we created an identical predictive model in IBM SPSS modeler to export the predictive model in PMML format. This model is in similar agreement of WEKA predictive model in terms of accuracy, recall, specificity and precision values.

IBM SPSS modeler flows - *Weka flows replicated in IBM SPSS*

OAAI Academic Anaytics Predictive model - *PMML format*

Deploying the OAAI Predictive model

Testing Datasets - Marist Undergraduate Fall 2011, Cerritos 2012 Spring, Savannah State 2012 Spring, Redwoods 2012 Spring

The kettle transformation process for the testing and scoring datasets, remains the same. The only difference would be that calculation of partial grades were not required as we were performing our predictions during the semester. Hence we were receiving updated Gradebook and Sakai Events datasets 4 weeks / 8 weeks / 12 weeks into the course

Scoring Phase using WEKA Scoring plug-in in Kettle.

Scoring is a process where the live datasets are fed into the trained Open Academic Analytics Predictive model to produce predictions on student data based on their performances in the semester by comparing it with the metrics of the training datasets. Once the pre-processed dataset from the ETL is available we deploy the trained WEKA predictive model using WEKA scoring plugin for kettle and predict the outcomes based on the student performances ,denoted by the metrics calculated. The final outcome are appended to the pre-processed datasets.

The predictive models developed based on SMO, Logistic Regression provides us with the probabilities or percentage of confidence in the prediction outcome. We further classify it into Risk categories by selecting a range for each category

- HIGH RISK
- MEDIUM RISK
- LOW RISK
- NO RISK

Reporting Phase using Pentaho Report Designer tool.

Using a Pentaho Report Designer tool we create a template for report generation. The resultant template can be embedded into KETTLE using Pentaho reporting plugin for Kettle.

Please find a sample Videos of complete deployment of OAAI predictive model and Academic Alert Report generation process. The current videos includes Academic Alert Report generation for Cerritos College.

Academic Alert Report generation for Cerritos College Video- *OAAI - AAR.wmv*

Sample flows - Cerritos Academic Alert Report 3 generation process - *Cerritos AAR3.zip*

Releasing the Data-mining flows with Pentaho

Initially in our project grant proposal we had proposed releasing, under an open-license, a Sakai "Student Effort Data" (SED) API that will capture user activity data and expose it in a secure fashion for use by early alert analytics-based tools. When we performed pilots in various partner institutions we found that there was a lot of variability involved since the partner institutions had many customization included on the base Sakai version tailored appropriately to suit the institutional requirements. As a result of this variability we realized that developing an API based approach to capture the user activity satisfying a common extraction process would be a tedious feat since it requires code changes to be made on various existing instances.

On a closer feasibility analysis regarding achieving a standard and uniform extraction process we found that use of data mining approach with Pentaho would be the best way to accomplish the task circumventing the variability factor introduced as a result of Customization of Sakai. The data mining approach requires no code change while extracting user activity from the various existing instances of Sakai, thus helping us to create an instance independent yet streamlined process.

Related Resources

- *Log Events Descriptions - Sakai 2.4.x* provides a detailed list of all the events available in Sakai CMS.

Contact Information

OAAI Principal Investigators and Researchers

- Josh Baron, Senior Academic Technology Officer, Marist College.
  Email : josh.baron@marist.edu
- Eitel Lauria, Graduate Director, Information Systems, Marist College.
  Email: Eitel.Lauria@marist.edu
- Sandeep Markondiah Jayaprakash, Learning Analytics Specialist, Academic Technology and eLearning, Marist College
  Email : sandeep.jayaprakash1@marist.edu

Academic Technology & eLearning, Marist College.

Marist OAAI Student team

Past Members

- Vennirai Sundararaju, Graduate Student & Senior Technical Consultant(OAAI), Marist College
  Email : Vennirai.Sundararaju1@marist.edu
- Mallika Devireddy, Graduate Student & Senior Technical Consultant(OAAI), Marist College
  Email : Mallika.Devireddy1@marist.edu
- Krishnan Kottaiswamy, Graduate Student & Senior Technical Consultant(OAAI), Marist College
  Email : krishnan.kottaiswamy1@marist.edu
- Nagamani Jonnalagadda, Graduate Student & Senior Technical Consultant(OAAI), Marist College
  Email : nagamani.jonnalagadda1@marist.edu

Press Releases

- **Campus Technology Innovators Award 2013 : OAAI introduction video -**