Data Analytics in Evaluating Student Forum: A Study of Sampled Forum Data

Hsui-lin Winkler
hwinkler@pace.edu
Seidenberg School of Computer Science and Information Systems
Pace University
Pleasantville, NY 10570, USA

Abstract

This paper presents results and discusses issues using data analytics for analyzing student forums. Useful techniques such as frequency counting, categorization, and classification were found to provide insightful information when applied to temporal student forum data. The main constraints in applying data analytics are direct accessibility to data sources and confusing user interfaces in data collection.

Keywords: data analytics, student forum, online teaching and learning, evaluation

1. INTRODUCTION

Student forums are now commonly used in online learning as a supplement to classroom discussion. For example, the Discussion Board in the Blackboard system (http://www.blackboard.com) provides a forum feature that can be used for either group discussion or can be open to all to post ideas or formal documents. In the most recent version, it also allows instructors to grade the forum or the discussion contents.

The basic functionality of a student forum is similar to many public blogs in that it captures a participant's name, posted topic, content, and the time when the content is posted. In popular news public blogs such as New York Times (http://www.nytimes.com) or Wall Street Journal (http://www.wsj.com), the blogs are setup as 'comments' for a specific editorial article and for everyone to read, but require registration to post. In student forums, restrictions can be added such that only a group of students or the whole class can either view or post to the forum. A thread can be created to begin a new topic area. With this categorization feature, a student forum can be used either to address specific questions or to answer other's comments. It can also be used as an un-guided group project discussion. Unlike a public blog, the participants in a student forum are restricted to students in the class and can be directly linked to a grade book.

In all types of forums, it often requires a lot of time to 'read' through this semi-structured content. A forum is an extra tool for the instructor if it is used for monitoring class participation without grading involved. However, if a forum is used as an evaluation tool, or if participation and contributions are used as part of the class grade, then it is a formal assessment task.
This requires sorting through all contents and to applying certain rubrics to the evaluation. In their paper, Dringus and Ellis (2005) presented an excellent review of the incentives and the benefits of using data mining as a strategy for evaluating discussion forums. Many common issues were raised as to how data mining techniques can be applied to automate the evaluation process.

In this paper, we apply data storage, data mining, and visualization techniques, or

collectively the techniques of data analytics, to sampled student forums, and summarize how effective these techniques can be when used to evaluate student forums. In the study, an open-source data forum system (http://www.phorum.org), as explained by Welling and Thomson (2003), is used to import and store sampled forum data. We will review the forum data, the data analytics methods, and issues involved in the analysis.

2. Sampled Student Forums

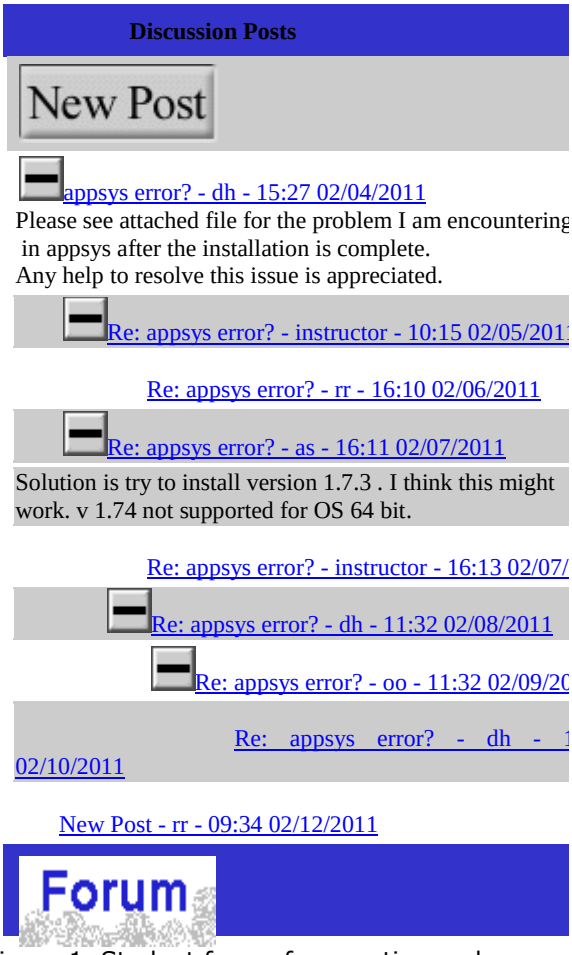

Figure 1. Student forum for question and answer.

When a student forum is part of an instructional system that has its own database to record all the data used in a specific class, the database is often not accessible by users. The only accessible data is through the user interface. Although the outcome of a student forum is readable and printable, it is too restricted to be used for data mining directly. In this study, we listed four different data samples retrieved from student forums and imported into an open forum such that the data stored in the testing database can be accessed and processed. Accessing a forum database can be one obstacle in using data analytics for student forums because access to the data is often not allowed at the database level.

We display below four different types of student forums.

General question/answer student forum. This is an easy and straightforward forum for students to post class related questions and for anyone to post help or answers. It is very useful for online learning, especially when software tools or implementation are involved. The case illustrated here is for students to use a student forum when installing a Web platform and development tool.

Figure 2. Student forum for instructor posted question and required student participation.

As captured in the student forum in Figure 1, some postings can help to resolve software tool installation problems. Also note that the display is a general tree structure with the initial posting aligned to the far-left and each answer posting following immediately below with an indent to the right. We intentionally do not show most of the contents or they are altered such that the data privacy is preserved.

Specific question/answer student forum. The second type of forum is similar to an exercise with the instructor posting a question and every student required to post an answer or comment. It is graded such that participation is mandatory and not optional.

As displayed in Figure 2, it has a simple 2-level tree structure. Often the answer can be an essay in style, therefore lengthy at times. It can be linked directly to a grade book, such that online grading can be recorded. However, it is time-consuming to conduct comparison among the students.

Student forum for class project. The third and fourth types of forums provide an effective way to monitor student participation in a project.

Figure 3(a). Student forum for class project, sample data 1 – even participation among all project members.

As shown in Figure 3(a) and (b), the two sample forums are to demonstrate the different project groups in student participation. Both projects have very good project outcomes, but one has an even distribution for all the four students in the team, the second one has one student essentially carrying out the project with help from one teammate. Without going through the detailed contents in the forum, simply judging from the data structure as displayed here, which contains the tree structure of the communication, the participant, and the event time, we have some glimpse of the collaboration of the group. However, if we want to analyze and quantify the collaboration effort in the group, data analytics tools are needed to process these seemingly simple yet information rich forum data.

Figure 3(b). Student forum for class project 2 – one student is the main contributor of the project.

There are two issues using a forum as a 'virtual meeting room' for a group project. First, with today's easy access to various social networks, students prefer to use more casual communication environments such as Facebook (http://www.facebook.com), Google (http://www.google.com), or email, instead of the more restricted Blackboard. Second, a general forum interface which allows a new forum to begin with little restriction can easily confuse the 'string' of the postings and make the classification difficult later. Nonetheless,

recorded forum data provides valuable insights into how digital collaboration works.

3. REVIEW OF DATA ANALYTICS - STRUCTURE VS. UNSTRUCTURED DATA

Data analytics techniques are commonly used today to provide information for organization decision-making. They can be grouped into two approaches according to how data was collected and stored - structured vs. unstructured data. However, the division can be superficial since there can be unstructured data contained in the structured contents. For example, most of the tabulated data can be considered structured, whether they are stored in a database or retrieved from a database and stored in tables. Within a table, if a data element contains a general description of certain feature, then it is an unstructured data element. On the other hand, we can easily have unstructured data such as a document that contains tables or tabulated reports that is well structured. Often we find that many of the data mining techniques are selectively used due to the specific domain of data collection and application. Furthermore, many of the image and video data are in general unstructured, unless they are well documented and stored.

Most structured data, especially those stored in databases, has a specific purpose embedded in the design which can later be queried to provide vital information for an organization's needs. Data mining techniques intend to apply to the basic data to provide new information aside from the routine reports. In doing so, the data mining techniques are often applied to a reorganized database such as a data warehouse or specifically collected data. Following Marakas's summary (2002), we review a few commonly used techniques below and point out how they can be applied to discussion forums.

Frequency Analysis - This is simply counting the number of certain data or text of interest that are present in the data. Most database query tools can provide this analysis. This analysis can be applied to forum data by counting the number of student participations either by initializing a discussion or by adding comments to other's posting. It is also used for word counting, although the word counting result needs to be interpreted.

Categorization – A feature used to group the same types of data for certain information, it can be embedded in a question or viewpoint. A simple way to do this in a forum is to begin a thread with a clear agenda that can later be used to categorize the discussion topic. However, we found students often would begin a new thread when posting an answer due to the confusing or imperfect design of user interface.

Association – This feature intends to do the basic data sorting for transactional data on the assumption that data entered together has a better chance of being related. It is commonly used in the so-called basket analysis, where a buyer of a laptop computer has good chance to also purchase a laptop carrier in the same shopping basket. There are different ways to implement this in a student forum. For example, we can associate a student who frequently initializes a new posting with one who tends to answer others more frequently as well. Whether this is true or not would depend on the objective of a student forum as well as each student's learning style.

Link – The connectivity of different attributes can perhaps indicate a pattern or certain trend emerging. In the forum shown in Figure 2, all students were asked to answer an instructor-posted question. There were six forums in the semester covering various subjects. We found that the most answered forum, or 100% students participated, is the social network related question. Using 'link' we can not only find the most interested topic in one class, we can also find the most interested topic for each student.

Sequence – The order by which events occurred. In forum data, a tree structure with time-based construct is an excellent sequence example.

Predictive Modeling – A simple model that can be used to predict what might happen next. For example, a curve fitting of collected data is often used as a theory for forecasting. In forum data, various collections of data can be built into a model, although its usefulness requires further testing.

Unstructured Data Analysis – Text mining is the general technique to search for pattern or extract for meaning in natural language written texts. An emerging interest in analyzing open forum data is to use a combination of certain simple tools including word counts and key word searches to conduct 'sentiment analysis' in blogs

or micro-blogs. The assumption is that sentiment can be extracted based on a large number of data. The use of positive or negative phrases can indicate the sentiment or reaction to certain events. It intends to provide a collective reflection. Leskovec (2011) has a review on how this technique has advanced in recent years. In student forums, this tool is not yet mature enough to replace traditional reading, marking, or applying pre-defined rubrics for evaluation. We still suggest the traditional 'reading' and apply pre-defined rubrics for evaluating the posting contents.

4. DATA ANALYTICS FOR STUDENT FORUM

In applying data mining technology, we often see various combinations of the above techniques. For example, frequency analysis can be used to categorize consumers who have purchased a specific product. Sequence can be used in conjunction with association to identify patterns of consumer behavior. How often a consumer who purchased product A also purchased product B in the same order or in sequence can provide valuable information to product management.

Our metrics used in evaluating or assessing student participation and learning outcomes in online forums are mostly derived from the face-to-face classroom meeting. Thus, we intend to establish a process using data analytics to show how the digital forum can provide similar measures that we think an effective forum should have.

An ordinary classroom discussion would include a question posted by the instructor or student and a string of answers from either instructor or students. Some discussions were graded during the class. Many were open-ended without recording other than some notes taken in the classroom. In a digital forum, similar criteria can be established. However, an online forum is often continuing for many weeks in a semester. It can consume a lot of time and effort to sort through all the records. Part of the reason why it takes a lot of effort to sort through the discussion board is because in a face-to-face situation we have more than 'reading' through the texts during the conversation. Our visual processes can place us back in the context of the discussion and help to more effectively review the forum.

In the following we use data from two project forums to demonstrate some of the methods in data analytics discussed above that can help to add context as well as content analysis.

4.1 Database Schema of Student Forum

Since the forum data collected contains content posted with the name of the participant, topic title, and time, it is straightforward to have metadata defined by these attributes. The only extra data needed is to identify whether the content of the message is a new posting or a reply to another existing posting. Here we used the design similar to the one specified in the open forum of phorum.org and instructed by Welling and Thomson (2003). The extra attribute needed for evaluation purposes is to have a grade or an equivalent quality indicator. The database consists of two tables: Heading and Body. Heading is used to define the data structure and has attributes of ParentID, ChildID, Title, Name, SubjectID, Time and ForumID. Body is defined by ForumID, Content and Score. The database as well as the user interfaces have been implemented and tested successfully.

4.2 Forum Participation Analysis

By simply plotting the number of postings recorded during the semester projects, we were able to 'observe' how students work within a group and through a semester. In Figures 4(a) and (b), posting number vs. time plots are shown for the overall participation in two project groups, for which a few forum data were shown in Figure 3(a) and (b). For group 1, 32 postings were recorded, but there seemed to be little recorded in the forum before April 7. The instructor later realized that it is because the group preferred to use a non-open forum for their early communication. If we compare the forum distribution with group 2 in Figure 4(b), in which 43 postings were recorded, it appeared that group 2 had more active participation in their project by this simple metric of temporal activity. However, if we look into the contributors, the distribution will reveal a different story.

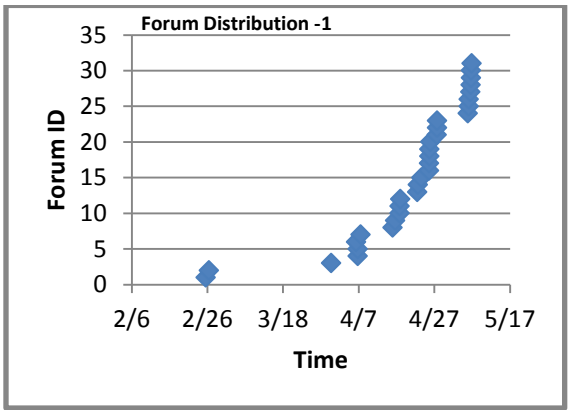

Figure 4(a) Forum number vs. time plot for project group 1.

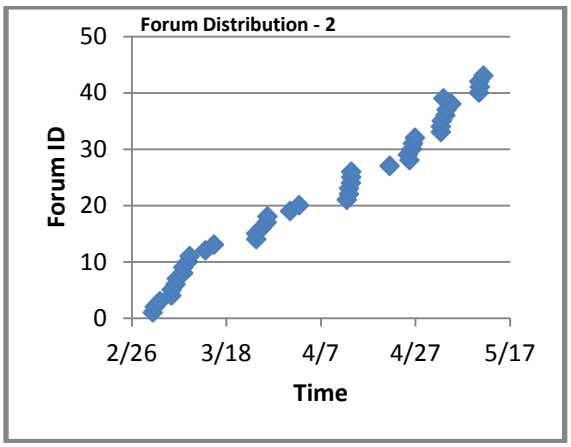

Figure 4(b) Forum number vs. time plot for group 2.

If we normalize a participant's contribution by calculating the percentage of each contribution with respect to the total, we can compare the two groups in Figure 5. Group 1 has roughly even contribution from each member, indexed here as pa, pr, yo and vi, whereas group 2 has half of their members (ba and ji) not using the forum much. The participant jo is the predominant contributor and el has participated in the forum at about the level as in group 1.

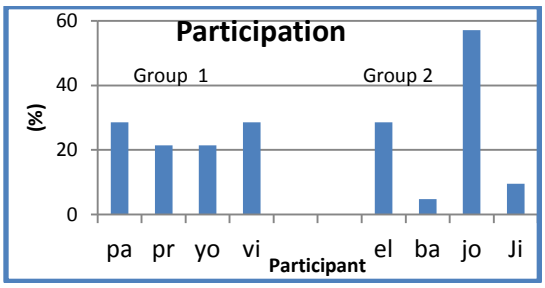

Figure 5. Normalized forum participation in the two project groups.

With the extra grade attribute added to the data system, it is straightforward to place students in different clusters using various cross plots. In Figure 6, we classify all students in their use of forum vs. the grade received. Three clusters can be drawn in this case.

In this study, we didn't intend to build a predictive model. Figures 4(a), (b) and Figure 6 can each be fit with a numerical curve that can further be used to forecast student participation in future project group.

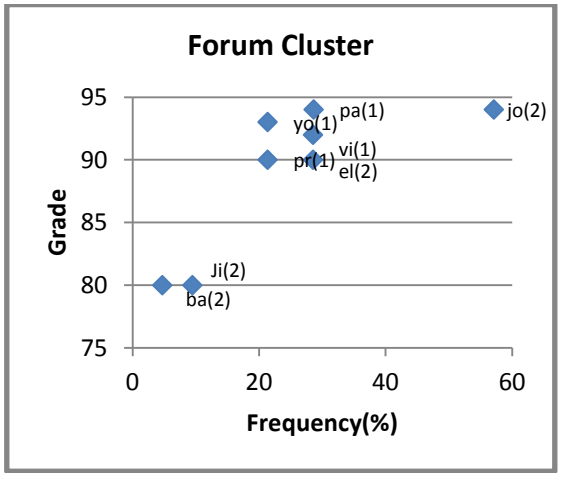

Figure 6. A cluster plot of forum grade vs. normalized frequency.

In this study, we have used SAS/BASIC in Enterprise Data Mining package (2004) for data processing. Some of the processing can also be implemented using an Excel Macro.

4.3 Forum Visualization

Most forums provide 'raw' data to viewers, or a direct display of a participant's name (or alias), link, message, and the posted date and time as shown in Figures 1-3. In this study, we have added the grade to replace the posted content in order to quantify the content quality.

Agrawala, Li and Berthouzoz (2011) advocated in their recent publication that the purpose of a visual design is to convey the viewer's perception and cognition of the underlying information. It is interesting to note that most asynchronous forums still use the same visualization as in the mid-eighties - texts embedded in a tree-structure as pointed out by Trampu and Grobelnik (2010). With the proliferation of mobile apps, we can expect to

see more effective visual tools that can aid the viewer of forums with quality assessment. An example of these is described by Engdahl, Köksal and Marsden in their mobile device application. (2004)

We plotted temporal participation to visualize the forum context as well as the collaboration efforts. However the visualization of an overall participation and temporal pattern of the forum remains to be implemented.

5. DISCUSSION AND SUMMARY

Student forums are becoming a popular tool to record various types of exchanges in an online classroom, yet an effective evaluation for the forum remains to be implemented. This study used data mining tools and an open forum database to show how data analytics can help to deliver more effective and quantitative evaluations.

Technology barriers remain in the accessibility of the database and in having a good user interface to guide the participant to post the message in the correct link. These issues are similar to those most mentioned in using data analytics. The difficulties are usually not in using of algorithm in data mining, but in the data itself. Often it takes significant time to 'clean' the data before they can be analyzed.

We have demonstrated in this paper that the simplicity of student forum data structures is amenable to analysis using data analytics. The results can provide more effective evaluation of student participation. In this study we combined open forum data structure and used commercially available analysis tools.

Future work will involve embedding the analysis tools and user interface in the open forum, and making them available to the open source community.

7. REFERENCES

Agrawala, M., Li W., & Berthouzoz F. (2011). Design Principles for Visual Communication, *Communications of the ACM, April 2011, 54 (4), pp. 60-69.*

Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing *asynchronous* discussion forums.
In Computers & Education 45 (2005) 141–160 Elsevier Ltd.

Engdahl, B., Köksal, M. & Marsden, G. (2005). Using Treemaps to Visualize Threaded Discussion Forums on PDAs, *CHI 2005,* April 2–7, 2004, Portland, Oregon, USA., ACM 1-59593-002-7/05/0004.

Leskovec, J. (2011). Social Media Analytics: Tracking, Modeling and Predicting the Flow of Information through Networks, WWW 2011, March 28–April 1, 2011, *Hyderabad*, India. ACM 978-1-4503-0637-9/11/03.

Marakas, G. M. (2002). Modern Data Warehousing, Mining, and Visualization: Core Concepts, Prentice Hall.

SAS Enterprise Miner, Data Mining Using SAS Enterprise Miner – A Case Study Approach. (2003) Copyright © 2003, SAS Institute Inc., Cary, NC, USA, ISBN 1-59047-395-7

Trampu, M., & Grobelnik, M. (2010). Visualization of Online Discussion Forums, In Editors: T. Diethe, N. Cristianini, & J. Shawe-Taylor, JMLR: Workshop and Conference Proceedings 11 (2010) 134-141 Workshop on Applications of Pattern Analysis.

Welling, L. & Thomson L. (2004). PHP and MYSQL Web Development, fourth Edition, Addison-Wesley.