Where Students Go and How They Do: Wi-Fi Location Data Versus Academic Performance

Soumabha Sarkar*, Bryan Carpenter*, Mohamed Bader-El-Den* and Arron Knight†
*School of Computing
University of Portsmouth
Portsmouth PO1 3HE, UK
Email: Soumabha.Sarkar@myport.ac.uk,Bryan.Carpenter@port.ac.uk,Mohamed.Bader@port.ac.uk
†Information Services
University of Portsmouth
Email: Arron.Knight@port.ac.uk

Abstract—Arguably students’ academic performance is one of the key areas that a university should focus on and perform further analysis on available data. One of the primary objectives of this research is to analyse the Wi-Fi AMP (Access Module Processor) data gathered by the Information Services Department of University of Portsmouth and determine how students are spending their time on university campus. Later these data have been used to examine if this information can be related to the academic performance of the student. With the analysed Wi-Fi data and academic records of the students a preliminary model has been created which tries to show how the time spent by the students in different types of location are correlated.

I. INTRODUCTION

It is now hard to envisage life without the Internet. The arrival of Wi-Fi moreover has made it easier for users to gain access to the Internet with laptops or other mobile devices connected to nearby Access Points. In the University of Portsmouth, for example, wireless access is available to staff and students throughout the university buildings and in over seventy other locations in the Portsmouth area. It can be accessed using a university user name and password. Portsmouth is part of the Eduroam community. Eduroam (education roaming) is a secure, world-wide roaming access service developed for the international research and education community. It allows students, researchers and staffs from participating institutions to obtain Internet connectivity across campus and when visiting other participating institutions.

Presumably students’ academic performance is one of the key areas of interest to any university. Students are the main asset for a university. So to be successful, the main focus should be on students’ need. Administrators in higher education are striving to achieve greater quality with fewer resources. Besides, not only are universities required to measure annual progress for students, but government (through ministries for education) aid is directly linked to these results. Hence, ability to analyse a student’s performance is very important in educational environment. A university needs data to do this kind of analysis to improve the quality of education. But analysing and reporting requirements across a university system is challenging. A way to overcome this challenge is to use the techniques of data mining. Data mining is the field of discovering potentially useful information from a large amount of data. This technique is widely used in retail business and bio-informatics.

Traditionally, educational researchers have used methods such as surveys, interviews, focus groups, and classroom activities to collect data related to students performance. These methods are usually very time-consuming, thus cannot be duplicated or repeated with high frequency [7]. But the pervasive popularity of the Internet in the past decade has largely changed the way people process information around the world [6]. Since the beginning of the Eduroam initiative in 2003, there has been a significant growth in university campus wireless networks all around Europe [4]. With the mass deployment of Wi-Fi, short for “Wireless Fidelity”, the usefulness of Wi-Fi network is no longer limited to connecting attached hosts together or connecting them to the Internet [11].

With the incorporation of Wi-Fi interfaces into many flavours of standard consumer hardware like mobile phones, laptops, personal digital assistants and portable media players, Wi-Fi positioning systems can be an excellent option to track the position of students within a university campus. In this paper we will correlate such tracking information with students’ academic performance to implement a new kind of Educational Data Mining (EDM). This brings together two types of data: students’ study related data and students’ location history in the university campus.

II. RELATED WORK

A lot of research has been done on Educational Data Mining (EDM) in the field of e-learning. In [19], a decision making system have been proposed in e-learning platform which helps distance instructors to know who their students are, how they work, how they use the virtual course, where they find the problems and so on, and in this way, instructors can act, for example, proposing new tasks, reorganising the content-pages, adding new information, opening discussions as soon as they detect any difficulty. Similarly, work has been done examining the influence of fourteen usage variables and the final grades of the participants of three large blended learning courses in order to learn about the dependencies between Learning Management System (LMS) usage patterns and learning results [16]. In their study, a good correlation was found for the students’ online activities within the e-learning platform such as solving examples in the selected course.
Apart from the implementation of EDM in e-learning, research has been done into attrition issues whereby students fail to progress through their course due to poor performance, by analysing the data collected from freshman- and sophomore-level programming courses [8]. On the other hand, researchers have introduced a framework for identifying at-risk students combining simple metrics, gathered from social network and statistical analysis domains, that have been shown to correlate with student performance and required slow manual data collection or additional expert analysis [10]. Also, the feasibility of using students’ electronic portfolio have been investigated as a resource for predicting student retention with help of a range of classification methods and their experiment showed significant accuracy to identify at-risk students at very early stages of their academic life, giving educators the opportunity to intervene in a more timely and effective fashion [1]. Some researchers have proposed a prediction model based on the Radial Basis Function (RBF) [2]. Its aim was to predict marks obtained by students in a subject and based on the results of predicted performance, students are categorised into groups and the students likely to fail are warned beforehand for improvement. Apart from that, an experiment was conducted by researchers to build an intervention prediction models focusing on rural schools in a developing country. The results from the experiment demonstrate that it is possible to attain reasonably accurate intervention prediction models even with a reduced dataset [15]. It has been observed in a research study that a collaborative multi-regression model provide better accuracy in predicting student performance than single linear regression as it consider personal student differences [9].

While many researchers have concentrated on predicting students’ performance, it has been examined how academic performance of students evolves over time during a study programme by analysing two consecutive cohorts using X-means clustering algorithm [3]. Similarly, the effect of different variables collected from the activity log on the final score of the student have been examined [5]. Also, now with the rise of cyber infrastructure, universities are providing Internet facilities in libraries and other open access areas to take the advantage of Internet’s global reach. Hence, the task of managing information technology services on university campus has been changed dramatically that paves the way for Web mining. Web mining in the domain of education can be categorised to three different category: web content mining, web usage mining and web structure mining [12]. Researchers have demonstrated how web usage mining methods have been taken up to address the challenges of understanding users and usage in higher education using cluster analysis and association mining [13].

The objective of prediction is to estimate the unknown value of a variable that describes the student. In education, the variable normally predicted is score or mark. The value of this variable can be numerical and continuous value or categorical and discrete value [17]. Studying the location of students in large buildings or groups of buildings within university campus raises new challenges on how to observe the motion: how to collect the student’s location inside buildings, how to observe a large number of students, how to do it in the entire building and how to collect data for long time periods. Problem in using the Wi-Fi logs for observing human motion is that each record on the log refers to a Basic Service Set (BSS) by the MAC address of the access point (AP), but do not include the physical location of that AP. Therefore, additional information is required about the physical position of each AP, to enable analysis processes. A Wi-Fi log includes a large number of parameters. For locating a device and its movement, the following parameters value have to be extracted: Timestamp, session ID, client device’s MAC address, Access Point’s MAC address, event type i.e. Start or Stop, and the name of the AP [14].

III. BACKGROUND AND METHODOLOGY

This project began when we were made aware that the University’s Information Services department had been collecting data on all Wi-Fi connections (connect time, connect location, user id, and so on) over an extended period. They were looking for opportunities for exploiting the large body of data collected. As academics we were naturally drawn to the possibility of using this data to explore student movements, and in particular to try to relate their whereabouts over some period of time to how well they do academically. An obvious thought is that one might ultimately be able to make evidence-based recommendations to students about where best to spend their time (library, lectures, etc) to improve their grades.

In this project, time spent by students has been measured with help of the information fetched from the Wi-Fi AMP (Access Module Processor) log files. The Wi-Fi data contains the studentid and the login information related to Wi-Fi enabled devices they are using with in the University campus.

The Wi-Fi Data is provided by the university information services department. The Wi-Fi data includes the information recorded in Table I. The Wi-Fi data are analysed to determine where and when the students were spending their time within university campus.

The second set of data are provided by the academic supervisor of the project. In this pilot study, only first year undergraduate students in School of Computing are taken into considered. This data set includes the early stage academic performance information of those students on two units taken by the majority students in the school, and also timetable information for the all students relating to those units.

Clearly in a project like this an overriding consideration is student privacy (and indeed privacy of all members of the university, since the original Wi-Fi data includes information potentially relating to all of them). The aim of the project is certainly only to release aggregate information to the outside world—never any information that could be related to an individual. Our policy beyond that was to restrict the flow of sensitive information even between collaborators within the project, based on anonymising identifying information through hash functions. So inevitably the network specialist (Knight) had access to the unobfuscated Wi-Fi data and the academics (Carpenter, Bader) had access to the unobfuscated academic data. But neither of these would see the others data (before anonymisation) and the principal researcher (Sarkar) would see no unobfuscated data. Moreover the anonymisation scheme depended on a secret key shared by the network specialist and the academics, but not known to the principal researcher. Finally the principal researcher signed a non-disclosure agreement restricting him from abusing the data.
The anonymisation scheme consisted of hashing the user-name information—considered as personally identifying information (PII)—in the Wi-Fi usage log data and in the academic data. In addition the MAC address of user’s device in the Wi-Fi data is hashed because this is also considered sensitive. The one-way hashing algorithm used was SHA256. To incorporate the secret key referred to above we actually used the HMAC-SHA256 algorithm, a widely implemented function from the networking field that conveniently combines the hashing algorithm with key information (the secret key prevents dictionary attacks by the principal researcher, in the presumably unlikely event he should violate the non-disclosure).

The anonymised data were then processed to get the information regarding the time spent by students in different buildings. The Wi-Fi log data contained thirty-one different types of information. Out of those, only four types of information were used to get the desired output.

- Device Name
- Username
- Connect Time
- Disconnect Time

Here device name indicates the name of the Wi-Fi Access Points through which user access internet by connecting their Wi-Fi enabled device. While installing the network, a well structured naming convention was implemented. All the access point in a particular building had unique names, but the first four characters of all access point in a particular building had always same. Also, it was possible to locate the floor of the building where the access point was installed as the sixth character of the access point name included the floor information. A sample of Wi-Fi AMP log data has been shown in Figure 1. Here LIBR-1-005-225 is the name of the access point or device name installed in university library first floor. In fact, all the access points installed in library starts with the prefix LIBR.

From sample in Figure 1, it can be inferred that the student with user-name up754977 spent around two and half hours in the library on 20th April, 2015. To extract this kind of information, Wi-Fi AMP log files as well as academic files for calculating average marks were kept in Hadoop File System (HDFS) and MapReduce job was run on Hadoop cluster.

For this pilot project, Wi-Fi AMP log files provided by the IS department covered slightly less than one week of activity. The size of the log was somewhat over 400MB. Ideally we would like to process data for at least a full academic year, which would be correspondingly much larger. That was one reason HDFS was used to store and process data. Initially a Hadoop Pseudo Distributed Environment, also known as Single Node Setup was created, which was a virtual machine that ran all of the Hadoop daemons simultaneously on a single machine. This allowed to quickly write scripts and test them on limited data sets without having to connect to a remote cluster. Java was used to write the MapReduce jobs which were run in Hadoop cluster for data processing as well as to develop the GUI of the application.

Students’ time table information was included to determine the time spent lectures more accurately. The time-table information were fetched using the unit register files. The Names of these register files follow the pattern UNIT-Group-type-Day-StartTime-Room-register.csv, where UNIT is the name of the unit; type is "lab", "lecture" or "tut" (for tutorial); Day is day of week; StartTime is when session starts; Room is room no, building is prefix of Room (e.g. LG-1.07 means Lions Gate Building room no 1.07). These register files contained only anonymised StudentID. Also, for this research, lecture time considered in this study included formal class room lectures as well as lab session. This new timetable information was included in the MapReduce job ran on WiFi data.

### IV. SOFTWARE CONSIDERATIONS

To get compact information regarding the time spent by the students on university campus from the Wi-Fi log data, a MapReduce job was run on Hadoop cluster. Separate directories were created on Hadoop cluster called UniWifiData to store the Wi-Fi AMP log files, AcademicData to store the academic records and TimeTableData to store time-table information. The MapReduce job was run by executing a JAR file on Hadoop cluster.

#### A. Time-Table Data Processing

Four generics of the mapper class for Wi-Fi data processing were

- LongWritable (input key type)
- Text (input value type)
- Text (output key type)
- Text (output value type)
The input to the mapper depends on what InputFormat is used. The InputFormat is responsible for reading the incoming data and shaping it into whatever format the Mapper expects. The default InputFormat is TextInputFormat, which produces the input key/value pair as LongWritable/Text. Here the input files were the unit register files and InputFormat was TextInputFormat. Hence, the LongWritable input key represented the offset location of the current line being read from the Input Split of the Unit Register files and the Text represented the actual current line itself. Default InputFormat was used here as input key was not significant in this project perspective.

Four generics of the reducer class for Wi-Fi data processing were:

- Text (input key type)
- Text (input value type)
- Text (output key type)
- Text (output value type)

To generate the student time-table, the mapper class fetched the anonymised studentID and the file name that contained it. The studentID is passed as key to the reducer. And from the file name, the Day part, Start-Time and Room information were passed as value. The reducer, grabbed each studentID and merged all the value for that particular studentID. The final output file contains the studentID and the merged time-table information.

B. Academic Data Processing

Similarly, the MapReduce job running for the Academic data produced a text file that contained the anonymised user name and a numeric value that represented the average marked obtained in exam.

C. WiFi Log Data Processing

The generics of mapper and reducer for Wi-Fi data processing were similar to that of Time-Table data processing. To meet the requirement of the MapReduce job i.e. calculate the time spent by a student in individual buildings, the mapper grabbed four information from input value: username, device name which could represent the building name, connect time and disconnect time. Then mapper class emitted the composite key of username followed by colon followed by first four character of the Device name or Access Point name which actually indicate the university building and a composite output value of the connect time followed by a comma followed by disconnect time. In this case, the mapper only included the usernames that were present in the output file produced by the MapReduce job run on academic files.

The output key/value from the mapper class was the input key/value of the reducer class (Figure 2). The Hadoop infrastructure performed a sort and merge operation on all those key-value pairs generated by all the calls to map() to produce a set of one or more partitions. When a call to reduce was made in reducer class, it was made with all the values for a given key i.e. in this case a combination of particular username and location which was in this case the university building name.

To generate the student time-table, the mapper class fetched the anonymised studentID and the file name that contained it. The studentID is passed as key to the reducer. And from the file name, the Day part, Start-Time and Room information were passed as value. The reducer, grabbed each studentID and merged all the value for that particular studentID. The final output file contains the studentID and the merged time-table information.

D. Application User Interface

Finally two output files generated by the Wi-Fi and academic MapReduce job were used to produce a scatter plot and regression line and equation that represented how spending time in a particular location on university campus might be related to the academic performance. One part of the UI took the location information i.e. the building name from the user and fetched how much time was spent by the user in that location from output file generated by running MapReduce.
Fig. 4. Data Processing Pipeline

job on Wi-Fi data. The application gathered these information for all the users listed in the output file generated by running the MapReduce job on academic data. It also took the average mark obtained by the user from this same output file. This time information and academic marks were stored in a table which is later used to generate a scatter plot and draw a linear regression line with help of Weka (Figure 4). This application built a regression model that looked like:

\[ M = c_0 + c_1 T_{\text{Library}} + c_2 T_{\text{Lecture}} + c_3 T_{\text{OpenAccessArea}} \]  

(1)

where \( T_{\text{Library}} \) represents time spent on Library, similarly \( T_{\text{Lecture}} \) is time spent in scheduled teaching sessions, \( T_{\text{OpenAccessArea}} \) is time spent in open access areas (computer suites available to all students for use in their own time) and \( M \) stands for academic performance of the student. Here, \( c_0 \) is a constant term which is the mean response value when all predictor variables are set to zero. As zero setting for all predictors is a plausible scenario in this research study, \( c_0 \) serves as a garbage collector for any bias that is not accounted for by the terms in the model. For this research study, as only students from school of computing were taken into consideration and time spent in open access area has been considered as time spent in Anglesea, Portland and Park Building.

V. RESULTS AND EVALUATION

A. Results

In an early proof of concept phase of this project, academic data for twenty-one MSc students in School of Computing was gathered for analysis. But out of twenty-one students, data for only twelve students were found in Wi-Fi AMP log files, because at that time we only had access to the logs for a period outside term time (clearly not ideal). Nevertheless this work, reported in [18], established the methodology and most of the software for the pilot study reported here.

As initial analysis did not include term time data, further analysis was done using a second set of Wi-Fi log data of four days during term time and academic records included early assessments of 349 first year undergraduate students from School of Computing. The academic records considered here included two units (on introductory programming and computer architecture) and time table information was fetched by running MapReduce job on the register files for those two modules.

To give a feel for the raw data prior to any model building or incorporation of timetable data, Figures 5, 6, 7 and 8 give scatter plots of academic performance (measured by early assessments on the two units) versus time spent in four relevant buildings on campus.

Adding timetable data and “semantic” interpretation of various building locations, the multivariate regression model described in Equation 1 for this data set gave:

\[ c_0 = 29.2398, \]  

(2)

\[ c_1 = 0.014, \]  

(3)

\[ c_2 = 0.1141, \]  

(4)

\[ c_3 = -0.0036 \]  

(5)

Test results from Ten fold cross validation and Percentage Split are shown in Table II. For comparison, models generated for
individual variables were

1) \[ M = 0.0397 T_{library} + 40.1649 \] (Figure 9)
   Correlation coefficient: 0.0707
   Mean absolute error (RAE): 26.4854
   Root mean squared error (RMSE): 30.2351
2) \[ M = 0.1139 T_{lecture} + 29.3556 \] (Figure 10)
   Correlation coefficient: 0.416
   Mean absolute error (RAE): 23.9133
   Root mean squared error (RMSE): 27.5527
3) \[ M = 0.0714 T_{openAccessArea} + 34.9312 \] (Figure 11)
   Correlation coefficient: 0.208
   Mean absolute error (RAE): 25.8627
   Root mean squared error (RMSE): 29.6464

It can be noticed that, although the combined model showed a negative coefficient for open access area, two variable regression lines showed a positive slope for library time, lecture time as well as time spent in open access area.

B. Evaluation

Bearing in mind that this pilot study involved only a few days location data and only two early assessments, we shouldn’t place a great deal of faith in quantitative results— their statistical significance is hard to assess. But it is interesting nevertheless to reflect on the results obtained so far.

<table>
<thead>
<tr>
<th></th>
<th>Ten Fold Cross Validation</th>
<th>Percentage Split with 66%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation coefficient</td>
<td>0.4042</td>
<td>0.4058</td>
</tr>
<tr>
<td>Mean absolute error (RAE)</td>
<td>24.0329</td>
<td>23.9896</td>
</tr>
<tr>
<td>Root mean squared error (RMSE)</td>
<td>27.7228</td>
<td>27.3589</td>
</tr>
</tbody>
</table>

TABLE II. ERROR VALUES OF THE REGRESSION MODEL

It was expected that all the coefficients of the predictor variable in Equation 1 would be positive and comparison would be done between those coefficients to find out which predictor impacts a lot on students performance and the generated model showed that lecture time and library time had a positive impact on student performance. As reported in equations 3 and 4, per unit time, lecture time had almost eight times more positive impact than that of library time, but in the period considered students would only have had at most 300 minutes of “lecture” time available to them related to the units considered (recall we included practical sessions here). With a coefficient in the model of 0.11, this would at face value suggest up to an average 30 point improvement in score if all available teaching time was taken up. We emphasise that

1 The time axis in Figure 10 extends beyond 300 minutes because time spent in a timetabled location immediately before or after a scheduled session has been included.
this interpretation is for illustration only—at this stage the data and models should be taken with a large pinch of salt.

In the model, time spent in open access areas had significantly less and slightly negative impact (equation 5). This is probably an artifact of fitting, interpretation, and low statistics, because the uni-variate fit to time in buildings with open access areas showed a small positive correlation.

While testing the regression model for the university buildings, it was found that the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are very high. In other words there is a huge amount of scatter, and models only cover average phenomena. In this pilot study, one fact that might affect the regression model was presence of outliers in the training data set. A regression model performs very badly when some points in the training data have excessively large or small values for the dependent variable compared to the rest of the training data.

VI. Conclusion

The research reported here can be considered a pilot study. The outcome of this study is useful to determine the student activity, department facilities that are important factors in terms of academic achievements. A great deal of work remains to be done to improve the accuracy of the model developed. In the ideal case, the research should take at least data for a whole academic year and all units taught in that year.

Beside greater data, the model needs refining. For example, in the current model, time spent by the students in Anglesea, Portland and Park building has been considered as time spent in open access areas because those buildings have such computer suites. Whereas actual time spent by students in open access area should be calculated by considering the specific access points in those areas. Also, one of the limitation of this research study is that it did not consider the fact that not all the students have Wi-Fi enabled devices and hence location data of those students are missing while creating the model.

In addition to that it has been mentioned that outlier variables affect the performance of Regression model. So, future work should take care of the issue. It should certainly examine other statistical modelling methods to build classifiers based on student performance.

VII. Acknowledgements

We are indebted to Andrew Minter and Tom Bosher of University of Portsmouth Information Services for inviting us into their early meetings on Wi-Fi analysis and facilitating this work throughout.

References


