ANLY482
Analytics Practicum
Project Proposal

Li Ka Shing Library Entry Data Analysis

Team 9 - Guardians of the library

Ren Mengxi | Wang Sijia | Wang Tianjing

Jan 2017

Table of Contents

Overview

- Sponsor introduction

Li Ka Shing Library (LKSLIB) is the first library of Singapore Management University, officially opened on 24 February 2006. The Library is named after Hong Kong businessman Dr. Li Ka-shing, and the Li Ka Shing Foundation donated and endowment to the library for collections. The main objective of the library is to offer an interactive study and research space for SMU community.

LKSLIB includes four floors that comprise about 8,800 square meters with 1,800 seats. Inside the library, there are a variety of spaces including open spaces for individual and collaborative use, learning commons which opens 24/7, quiet areas that for individuals to focus on their work, project rooms with LCD panels, investment studio, postgraduate lounges etc. As a modern library, it is also well equipped with high-speed wireless network, color printers, scanners, public computers with professional financial software available, up-to-date newspapers and magazines, collections of lifestyle videos and games, and this is also the reason why LKSLIB is so attractive for SMU community.

- Project introduction

In our project, our focus is on analyzing the library entry information from the card reader logs. The card readers are located at the entrance of the library, both located at the main entrance of LKSLIB and at the linkbridge side entrance, and students need to tap their card whenever they enter the library. This provides us with the entry information, which includes timestamp and basic information about the student. To better understand the library usage, the library management team is interested to know whether we could find a pattern about the usage of library for a particular user group (e.g. Information Systems undergraduate, Year 2), and if any business insights could be drawn from the data. Another part of the analysis is about group detection, meaning that several people always tend to appear together in the library, to see how could library provide better study environment for group users. New topics such as hogging rate analysis (together with the occupancy data of the library) may also be considered.

Motivation

The management team of LKSLIB is striving for better customer experience, especially the physical study environment in LKSLIB. To get more information from the library user, they have collected information from social media platforms, user surveys feedbacks, etc. However, unlike the information about online search request, it was hard for them to collect statistics about the physical usage of the library, especially the usage of a specific user group.

LKSLIB utilizes card reader to control the entrance of the library, and when student approaches the entrance and taps in, a log will automatically be generated with the student information and the timestamp. Recently, the library management teams decide to look into this dataset to have a better understanding about the physical usage of the library. Combined the card reader log with the student database, they could get the entrance information about each single library user, but this is not enough. With the help of data analytics technology, the management team is looking for more detailed insight about the students' usage of the physical library.

Objectives

The aim of our project is to help the library management team to have a deeper understanding about the student usage of the physical library with the help of the card reader data, and to support their decision making process of the improvement in user experience with data analytics technologies.

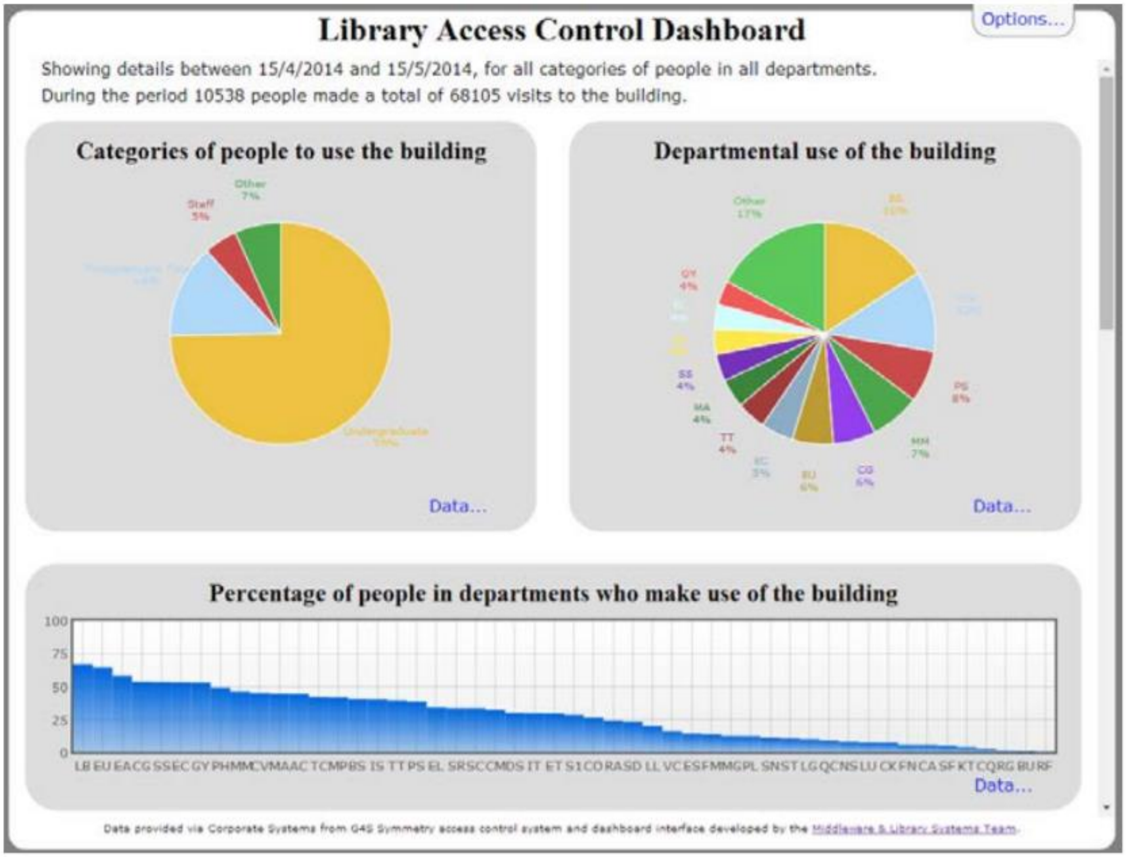
The objectives of our project consist the following:

- To summarize the library entrance information of a specific user group
- To detect the group of users who usually stick together
- To analyze the hogging situation in the library
- To build a tool for library management team to visualize the data collected interactively

Literature Research

1. Georgia State University (GSU) Library has required students to swipe their campus ID card to enter since 2002 for security control, and the data generated through card swiping could be collected [1]. Since January 2009, a new analytics system by Advanced Campus Services (ACS) has been implemented to make use of data generated by the system in order to help the library to provide better service. This system can generate reports on the total number of swipes and the number of swipes for unique visitors. Therefore, the user will be able to see how many unique students of certain admission year have visited the library during certain time period. Our group think we can apply the same logic into our project as well by analyzing the percentage of each type of students and how many times does each student enter the library with various breakdowns. Another point that GSU team analyzed is the address of the student (stay in the dorm or not). If we could get related information and further look into this aspect, we can draw insights about the relationship between the living area and the frequency the student comes to the library, and figure out whether students living nearer to school are tend to visit library more often.

2. There is a group of students from Loughborough University publishing a visualizing model to understand the usage of Pilkington Library. Their main focus is on categories of people who enter the library and departmental use of the building. The final presentation inspired us to include a dynamic dashboard in our final product which allows the management team to look at the breakdown by defining a certain time period.



Library Access Control Dashboard

Showing details between 15/4/2014 and 15/5/2014, for all categories of people in all departments.
During the period 10538 people made a total of 68105 visits to the building.

Data provided via Corporate Systems from G4S Symmetry access control system and dashboard interface developed by the Middleware & Library Systems Team.

Options...
Categories of people to use the building
Other 7%
Staff 5%
Data...
Departmental use of the building
Other 17%
QY 4%
SS 4%
MA 4%
TT 4%
IC 5%
BU 6%
CG 6%
MM 7%
PS 8%
Data...
Percentage of people in departments who make use of the building
100
75
50
25
0
LB EU EA CG SS EC GY PH MM CV MA AC TC MP BS IS TT PS EL SR SC CM DS IT ET S1 CO RA SD LL VC ES FM MG PL SN ST LG QC NS LU CK FN CA SF KT CQ RG BU RF
Data...

Figure 4: Screenshot of access control dashboard

(source:https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/15522/3/visualising-access-final.pdf)

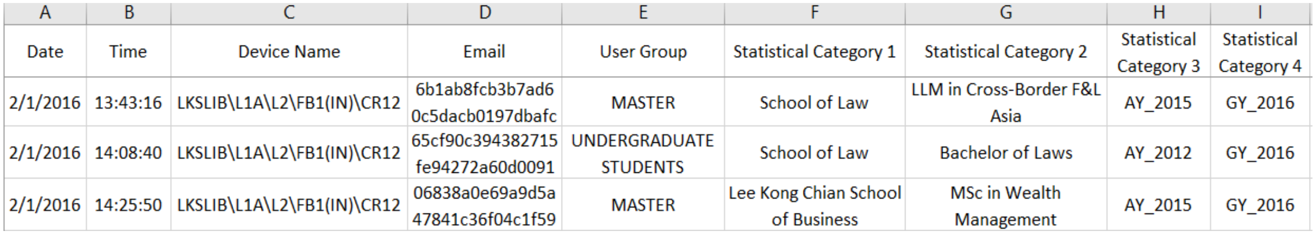
3. We also found a space utilization study of Princeton Theological Seminary Library [2]. Besides the gate counts, the research team also collected data like room reservation information, observational data and other qualitative data through surveys. This gives them advantages to have a broader view about the space utilization analysis. Although the paper does not give detailed analysis except key findings, it will be a good move forward direction of our study if library could cooperate with other departments to get related data all together in the future.

Data

The dataset is provided by library analytics and research team in csv format. As they are still in progress of matching the card entry logs with student school information, the data comes monthly by monthly. We will concatenate all the tables and study on one-year data of 2016.

The original dataset includes:

- Date: the date of the entry in format of d/m/y
- Time: the time of the entry in format of hh:mm:ss
- Device Name: gantry number (LKSLIB\L1A\L2\FB1(IN)\CR 12/14/16/18)
- Email: the hashed email address of the entered student
- User Group: which group the student is in (undergraduate/master/phd)
- Statistical Category 1: the school of the student (LKCSB/SOA/SOE/SIS/SOL/SOSS)
- Statistical Category 2: the major of the student
- Statistical Category 3: the admission year of the student
- Statistical Category 4: the graduation year of the student. If the student has not graduated yet, this field will be shown as 0.

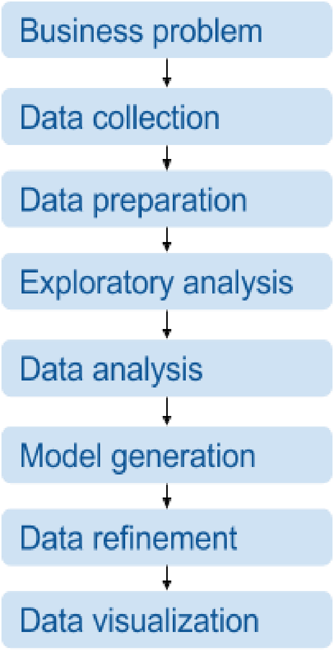


A	B	C	D	E	F	G	H	I
Date	Time	Device Name	Email	User Group	Statistical Category 1	Statistical Category 2	Statistical Category 3	Statistical Category 4
2/1/2016	13:43:16	LKSLIB\L1A\L2\FB1(IN)\CR12	6b1ab8fcb3b7ad60c5dacb0197dbafc	MASTER	School of Law	LLM in Cross-Border F&L Asia	AY_2015	GY_2016
2/1/2016	14:08:40	LKSLIB\L1A\L2\FB1(IN)\CR12	65cf90c394382715fe94272a60d0091	UNDERGRADUATE STUDENTS	School of Law	Bachelor of Laws	AY_2012	GY_2016
2/1/2016	14:25:50	LKSLIB\L1A\L2\FB1(IN)\CR12	06838a0e69a9d5a47841c36f04c1f59	MASTER	Lee Kong Chian School of Business	MSc in Wealth Management	AY_2015	GY_2016

Scope of Work

Our scope of work includes analyses the business problems and define the issues. In order to compile the necessary data, we need to prepare the data for exploratory analysis after collecting the data. Missing values, multiple entry for some users and outliers have been found in the record. Some column format need to be set up and renamed before we start to perform the analysis.

We will constantly ask feedbacks from the sponsors and our supervisors to clear doubts and make further improvement on our projects. We will deliver a web application to allow the users to explore through the entry data in an interactive and visualized way.



Data Preparation

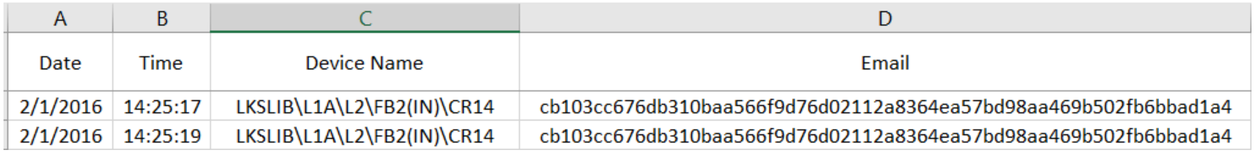
a. Variable transformation

Firstly, we will need to rename some columns and values. Also, as LKSLIB has different operation hours for weekday and weekends, it will be useful for us to include one more column to indicate the day of week. The revised columns will be: Date, Time, DoW, Gantry No., Email, User Group, School, Major, Admission Year, Graduation Year. The four different gantries will be mapped to A, B, C, D accordingly.

For the timestamp, it is not really meaningful to study until second level, so we decide to bin it. Currently our target is to bin to hourly because we are based on one-year data. Smaller bin basket may be considered if there is need to study the trend for a shorter period.

b. Multi entry data

Through observation we found that sometimes the same person has two records together, indicating that he entered library twice within a very short time frame, which is not practical.



A	B	C	D
Date	Time	Device Name	Email
2/1/2016	14:25:17	LKSLIB\L1A\L2\FB2(IN)\CR14	cb103cc676db310baa566f9d76d02112a8364ea57bd98aa469b502fb6bbad1a4
2/1/2016	14:25:19	LKSLIB\L1A\L2\FB2(IN)\CR14	cb103cc676db310baa566f9d76d02112a8364ea57bd98aa469b502fb6bbad1a4

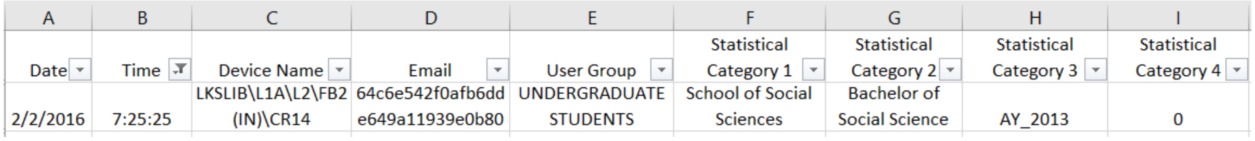
After discussing with our sponsor, we think this is caused by occasionally double tapping when someone tried to enter library. Therefore, we set a basic rule to clean up these multi entry data by removing the second entry if same email address appears twice within 10 seconds. This time frame is subjected to changes if we find any other reasons that will also cause the multi entry issue in the future.

c. Missing value

Some master students did not register their major on library system, which gives us 0s under Major column. Most of these students are from SIS, and few from LKCSB. As major is not possible to manipulate, we will discard these records at this moment.

d. Outlier

We also found one outlier for now in our dataset. There is one entry occurred out of library operation hour at 07:25AM. This entry will be discarded from all analysis.



A	B	C	D	E	F	G	H	I
Date	Time	Device Name	Email	User Group	Statistical Category 1	Statistical Category 2	Statistical Category 3	Statistical Category 4
2/2/2016	7:25:25	LKSLIB\L1A\L2\FB2 (IN)\CR14	64c6e542f0afb6dd e649a11939e0b80	UNDERGRADUATE STUDENTS	School of Social Sciences	Bachelor of Social Science	AY_2013	0

Methodology

With the data clean in hand, we will start off with an exploratory analysis into the behavior of library users and seek to understand the basic trend of library usage such as peak hour analysis and so on. This is mainly done by plotting the data points with different grouping combination using JMP and Tableau.

Our main focus will be on Clustering Analysis. We attempt to cluster the users into different groups using their categories, to observe whether students from same school have the habit to visit library around certain time, or what characteristic does a group own if they share similar pattern of usage.

Following by that, we will also conduct Association Rule Analysis, to detect whether students visit library individually or more as groups. If association relationship does exist, we will try to analyze what is the average group size and when is the most preferred time for group visiting. This will help library team to better allocate seats and other resources.

After above analysis, we may refine the user segments to have better differentiation between groups. From the results, we will generate recommendations and identify high usage students, as well as their preferences for LKSLIB. Lastly, if possible, we can extrapolate the data to forecast the demand of the library.

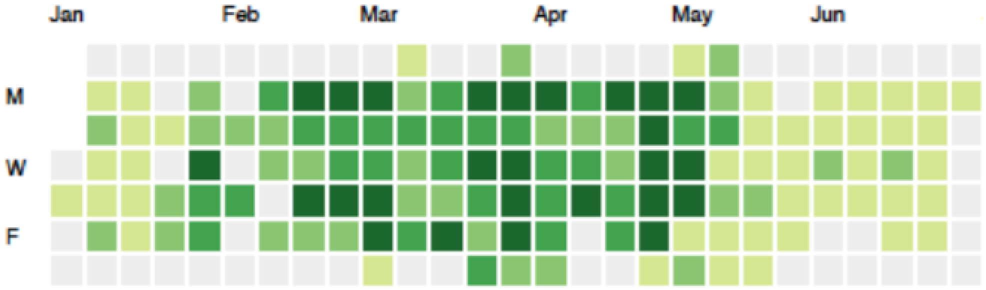
- Technologies
  Bootstrap, JMP, SAS, Tableau, Java, Sql, D3.js, JavaScript

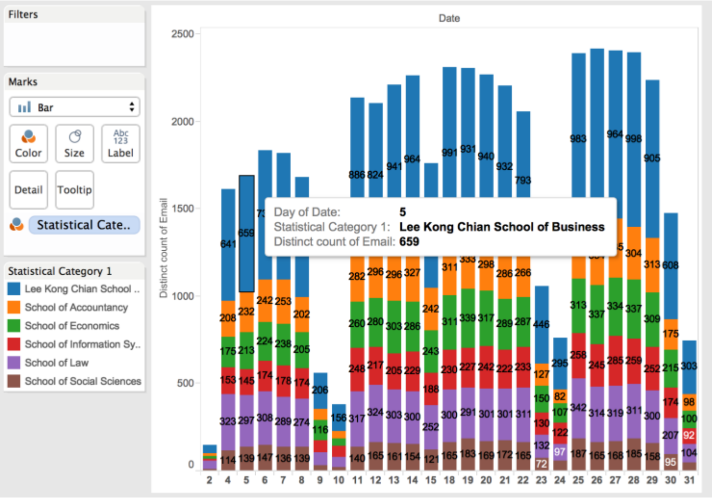
Deliverables

Through our project, a web application will be brought over. This web application would be able to allow the users to filter the date (year, month, day), time, faculty, user group of the LKS LIB entry data in an interactive way and show more visualized and more understandable diagrams and charts. By analyzing the generated diagrams, users can have a general understanding about the characteristics and behaviors of the specific user groups entering the library. It also potentially provides the librarians with better perspectives on certain issues currently exist in the library like the seat hogging. Furthermore, it may help them make decision on seat allocation in library to better serve the students community from different faculty in the future.

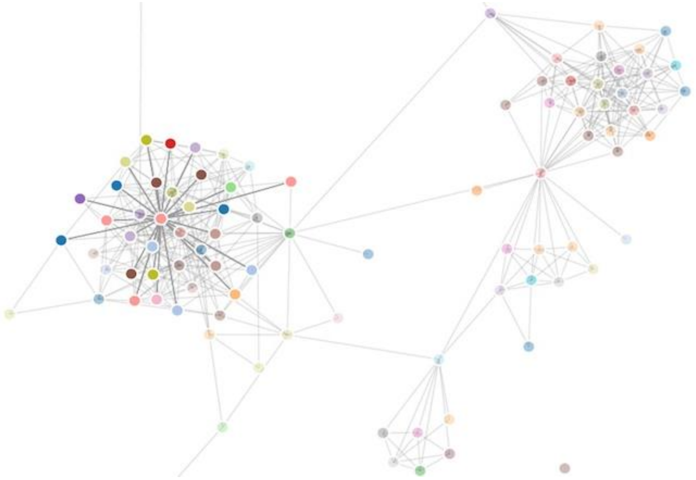
- Possible visualization:

Calendar view: providing the yearly view of the historical occupation rate





Day of Date: 5
Statistical Category 1: Lee Kong Chian School of Business
Distinct count of Email: 659



Bar Chart: visualizing the library entry data by different breakdowns

Social Network Diagram: showing the group detection results

Stakeholders

The primary stakeholders of this project are:

- Project Supervisor: Prof Kam Tin Seong, Associate Professor of Information Systems; Senior Advisor, SIS (Programme in Analytics)
- Sponsor:
  - Nursyeha Binte Yahaya, Learning and Information Services, Librarian
  - Aaron Tay, Manager, Library Analytics & Research Librarian Accountancy

Work Plan

	Week 1	Week2	Week3	Week4	Week5	Week6	Week7	Week8	Week9	Week10	Week11	Week12	Week13	Week14	Week15	Week16
Milestone Proposal - 1 Jan																
Project hunting	M+S+T															
Gathering Requirements/Data	M+S+T															
Ideas brainstorming	M+S+T															
Wiki page creation	M															
Tools installation and setup		M+S+T														
Proposal		M+S+T														
Data Cleaning		M														
Milestone Interim - 19th Feb																
Data Exploration		M+S+T														
Clustering analysis			M+T													
Assisociation analysis			S													
Data visulization					M+T											
Application development					M+S+T		Buffer									
Interim Report and Slides						M+S+T		Buffer								
Minutes record		S														
Wiki Update		T														
Milestone Final - 2nd April																
Gathering feedback								M+S+T	S							
Model refinement									M+S+T							
Testing										M+S+T						
Finalizing the system										M+S+T			Buffer			
Wiki Update									T							
Research paper													M+S+T			Buffer
Poster															T	
Presentation																M+S+T

Legend	
M - Mengxi	Coder
S - Sijia	BA
T - Tianjing	PM

Limitations and Future Work

Users may not be able to further explore the entry data by filtering more variables of the students (age, gender) due to the limited information of the user groups on the provided data. The work on joining of the library entry data and student profile database is still in progress so further development on implementing more functions to have better view on the student profiles can be looked into.

References

1. http://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1020&context=univ_lib_facpres
2. https://library.ptsem.edu/content/documents/LibrarySpaceUtilizationStudy_v3.pdf
3. http://mothership.sg/wp-content/uploads/2014/04/SMU_Library.jpg
4. https://library.smu.edu.sg/about-us/overview/about-us-li-ka-shing-library
5. https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/15522/3/visualising-access-final.pdf