

**ANLY482 Analytics Practicum**

**Project Proposal**

**(Li Ka Shing Library Proxy Log Analysis)**

**Team BJJ**

Members:

Lim Yu Xiang Bendexter

Tan Jun Rong

Wang Jing Xuan

Executive Summary

The Li Ka Shing Library’s electronic search platform offers a wide array of research resources with over 360,000 books, 80,000 journals, 160 databases and more than 16,000 SMU research publications in its Institutional Repository and Oral History Collection. The extensiveness of the amount of resources would mean nothing if the average user (e.g. You and me) do not utilize it to its fullest capabilities.

Therefore, the analytics team (part of Learning & Information Services) in Li Ka Shing Library would like to discover meaningful insights about user behaviour on its electronic resources to provide necessary assistance in forms of library e-resources training, helpdesk and support. However, the current problem is the lack of knowledge to handle the proxy log data collected from the library’s main web page, <http://library.smu.edu.sg/>. Thus, the proxy log data files are often neglected and not used at all.

We would like to realize the full potential of this data by first understanding the user behaviour of library’s electronic resources. After which we would aim to understand the relationship between different search queries and how it varies from certain clusters of users. Lastly, we would dive down to the details and examine the event sequence for unique users, in terms of how their search querying ‘journey’ appears.

Introduction

Sponsor Background

The Li Ka Shing Library, Singapore Management University’s library centred, was officially opened on 24 February 2006. The library was established and named after Hong Kong businessman Dr. Li Ka-shing, Chairman of Cheung Kong (Holdings) Limited and Hutchison Whampoa Limited. The Li Ka Shing Foundation also donated and endowed the library for collections and to Singapore Management University (SMU) for scholarship. The main purpose of the Li Ka Shing Library is to provide academic and professional knowledge resources and services to support the research and learning needs of the SMU community.

Today, the Li Ka Shing Library offers facilitation of knowledge creation via its electronic search platform and a wide array of research resources on and off campus. With over 360 000 printed and electronic books, over 80 000 printed and electronic journals, more than 160 electronic databases, over 16 000 SMU research publications in its Institutional Repository, and Oral History Collection, the Li Ka shing Library is a platform for the SMU community to enhance learning, both individually and collaboratively.

Minister Mentor Lee Kuan Yew has supported the Li Ka Shing Library to be seen as “the intellectual hub and a centre for research for faculty; as a place for students to come and collaborate.” In recognition of its effort towards improving effectiveness, productivity, and in building a culture of continuous improvement, the Li Ka Shing Library won the Outstanding Department Award at the Business Excellence Awards event hosted by President Prof. Arnoud De Meyer.

In essence, the Li Ka Shing Library is affectionately known to us students as the hub of the city campus where we spend most of our days revising and using both online and hard-copy Library resources.

Organization Problem & Motivation

The role of the analytics team (part of Learning & Information Services) in Li Ka Shing Library is to discover meaningful insights about user behaviour so as to provide necessary assistance in forms of library e-resources training, helpdesk and support. However, the current problem is that they do not know what to do with the logging data collected from the library’s main web page, <http://library.smu.edu.sg/>. Thus, the logging data files are neglected and therefore the library analytics team wishes to collaborate with us in realizing the full potential of this data.

Project Objectives

This project aims to do analysis on log files to:

1. Understand user behaviour by using a data-driven approach
2. Understand the relationship between different search queries for different users
3. Examine the event sequence for unique users (Eg. What articles did User A searched together or 1 after another in sequence)

Data Set Description

Preliminary Data Source

Proxy log data & student information data (Names of Students are Hashed)

Data Dictionary

Proxy Log Data

*59.189.71.33 tDU1zb0CaV2B8qZ 65ff93f70ca7ceaabcca62de3882ed1633bcd14ecdebbe95f9bd826bd68609ba [01/Jan/2016:00:01:39 +0800] "GET http://heinonline.org:80/HOL/VMTP?base=js&handle=hein.journals/fchlj23&div=7&collection=journals&input=(The%20Great%20Peace)&set\_as\_cursor=19&disp\_num=20&viewurl=SearchVolumeSOLR%3Finput%3D%2528The%2520Great%2520Peace%2529%26div%3D7%26f\_size%3D600%26num\_results%3D10%26handle%3Dhein.journals%252Ffchlj23%26collection%3Djournals%26set\_as\_cursor%3D19%26men\_tab%3Dsrchresults%26terms%3D%2528The%2520Great%2520Peace%2529 HTTP/1.1" 200 2121 "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36"*

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Description** | **Example** |
| Http address | This is the IP address of the webpage | 59.189.71.33 |
| Session ID | Each session is identified by an unique ID, which corresponds to 1 session by a single user | tDU1zb0CaV2B8qZ |
| Unique Student ID (Hashed) | The student ID is hashed by the SMU Library so as to protect the identity of users | 65ff93f70ca7ceaabcca62de3882ed1633bcd14ecdebbe95f9bd826bd68609ba |
| Timestamp | This is the timing which the log is recorded, and the log is recorded whenever the user performs a task. The time is in 24 hours format and in local Singapore time GST+0800. | [01/Jan/2016:00:01:39 +0800] |
| HTML method | The search query by the user typically comes after this HTML method. | GET |

Student Information Data *“feb0e4d05b236c0bcc0c7331dc754921cf9189c4c1317b0b112696fcf68cd2f8, MASTER School of Accountancy, MSc in CFO Leadership, AY\_2014, GY\_2015”*

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Description** | **Example** |
| Unique Student ID (Hashed) | This is provided so that we can match the unique student ID to the corresponding ones in the proxy data logs. | feb0e4d05b236c0bcc0c7331dc754921cf9189c4c1317b0b112696fcf68cd2f8 |
| Level of Education | This indicates which level of education the user is in, typically Masters or Bachelors programme. | MASTER |
| School | This indicates the school that the user is from. | School of Accountancy |
| Type of Programme | This indicates the specific programme the user is undertaking. | MSc in CFO Leadership |
| Admission Year | This indicates the year which the user is admitted into SMU. | AY\_2014 |
| Graduating Year | This indicates the year which the user is graduated from SMU. | GY\_2015 |

Work Methodology

Tools

* Trifacta for data wrangling
* Text Explorer on JMP Pro 13

Data Collection

* Ezproxy browser records each individual’s action in each session
* Each individual’s entry and exit from library is recorded

In complement to the datasets provided by the Li Ka Shing Library, we would also collect additional data such as:

* Dates of public holidays
* Period of semesters (1, 2, 3A, 3B)
* Periods in semesters where researches are mostly done (E.g. 2 weeks before project submission)
* Period of recess week
* User experience and User’s general behaviour through questionnaire

Data Preparation

As the data file is large (2.5Gb) and in .txt format, we will first find a software that is able to open such a file for us to even begin our data cleaning process. After which, as the log data consists of long strings with html tags, browser names and other parameters which may not be useful for the scope of this project, we will perform the appropriate techniques to break the strings up into separate tokens for better analysis.

Exploratory Data Analysis (EDA)

1. Descriptive analysis
2. Association analysis
3. Visualization of association
4. User behaviour analysis
5. User clustering

Risks and Limitations

The analysis of the proxy logs may prove to be a challenge as it is something unfamiliar to us; we need to explore various methods to pre-process the dataset and derive useful information from it. There is a probable risk that some search queries may be missing from the database log due to the design of the system, thus we may not be able to perform the analysis for all database resources that the SMU Library is currently subscribed with,

Key Stakeholders

Project Supervisor

Professor Kam Tin Seong, Associate Professor of Information Systems; Senior Advisor, SIS (Programme in Analytics)

Li Ka Shing Library

Analytics Team

* Aaron Tay, Manager, Library Analytics & Research Librarian
* Nursyeha Binte Yahaya, Librarian

Project Deliverables

For the Sponsor

At the end of this project, the following deliverables will be achieved and handed over to the Li Ka Shing Library Analytics Team:

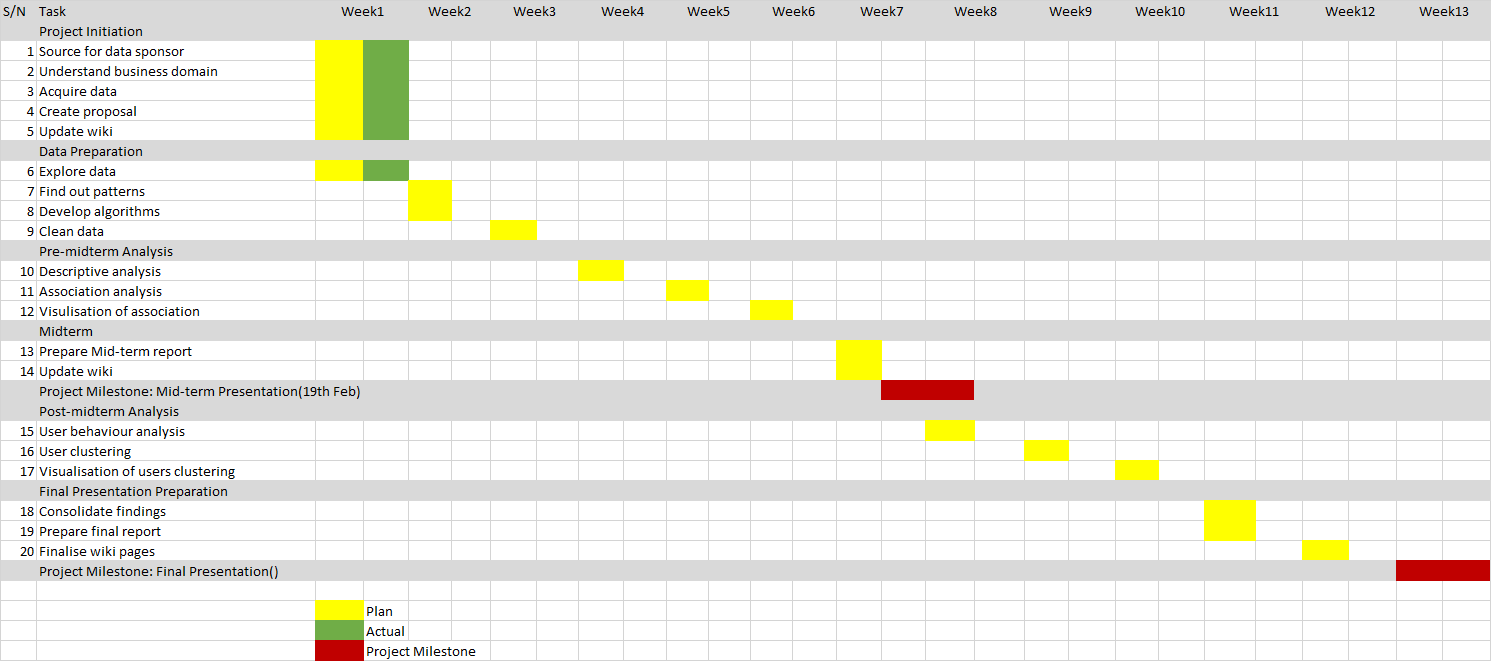
1. Report Analysis
   1. Methodology explanations to handling the given dataset
   2. Visualizations to portray event sequence for each unique users
   3. Text Analysis to explain users’ behaviours

For Singapore Management University

The following deliverables, in accordance with SMU ANLY482 Analytics Practicum requirements, will be achieved and submitted:

1. Project Proposal
2. Interim Presentation Slides
3. Interim report
4. Final Presentation
5. Final Report
6. Project Poster
7. Updated Wiki Page

Timeline



References

Singapore Management University. 2015, April. Li Ka Shing Library. Retrieved from <https://library.smu.edu.sg/about-us/overview/about-us-li-ka-shing-library>.

Singapore Management University. 2016, December. About Us – Overview. Retrieved from <https://library.smu.edu.sg/about-us-overview>.