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 History Collection. The extensiveness of the amount of resources, in terms of variety and quantity, would mean nothing if the average user (e.g. You and me) do not utilize it.[1]

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, there are a number of problems that we aim to resolve through these insights:

1. Lack of understanding on how users behave when navigating and searching through each database and/or e-resource
2. Lack of understanding on how different search queries are performed for different users
3. Lack of a proper search optimization engine for the search bar on library's main search page, thus unable to improve user experience in the different search queries for unique users

The proxy log data files are often neglected and not used at all. The Ezproxy log files has a size of up to 35GB, and are in text format. Based on the data set we have also derived the total volume of queries, as stated below:

1. Number of Daily queries: 236,446
2. Number of Weekly queries: 1,655,122
3. Number of Monthly queries: 6,620,491

The Li Ka Shing Library is currently attempting to understand the general user behaviour with regards to their resources – Number of student or staffs entering and exiting the library, the potential reach for each e-resource, and the impact of the Academic Writing module on these users' search techniques. They have been leveraging on their turnstile report to understanding how frequent students enter the library, and if there is a need to change the operating hours of the Library. In addition, they are looking into understanding the potential reach for each electronic database, but with their lack of knowledge in querying their dataset, they seek assistance to retrieve these insights. Hence, we would like to realize the full potential of this data by first understanding the user behaviour – based on their academic year, school, time, date, and other factors that we seek to discover throughout the course of this project. 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.

This project is a research-based undertaking, where we would present findings only. There will not be any additional resources built throughout the course of this project.

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. 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.[2]

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 to better discern the reach for each e-resource and the querying capabilities of each user category.
2. Understand the relationship between different search queries for different users
3. Examine the event sequence for unique users (E.g. What articles did User A searched together or 1 after another in sequence) to provide recommendations for

improvement in User Experience. For example, these event sequence insights could complement the optimization of the Search function, to suggest other searches that are based on each unique users' event sequence.

DATA SET DESCRIPTION

Preliminary Data Source

1. EzProxy log data
2. Student information data (Names of Students are Hashed)
3. Turnstile Report data

Data Dictionary

Proxy Log Data

Example of Raw 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"

Frequency count of Dataset: 6,620,491 (In a period of 1 month for Initial Dataset)
Name of Resource being looked up: https://library.smu.edu.sg/

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	[01/Jan/2016:00:01:39 +0800]

	and in local Singapore time GST+0800.	
HTML method	The search query by the user typically comes after this HTML method.	GET
Number of Results	This is the frequency count of the number of resources provided through that search keyword	%3D600%26num_results%3D10%
Search Query	This is the name of resource being looked up.	%2528The%2520Great%2520Peace%

Student Information Data

Example of Raw Data:

"feb0e4d05b236c0bcc0c7331dc754921cf9189c4c1317b0b112696fcf68cd2f8, MASTER School of Accountancy, MSc in CFO Leadership, AY_2014, GY_2015"

Frequency of count: 22,427

Name of Resource being looked up: SMU Internal Student Database (Matched based on email address)

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

Turnstile Report Data

"2/1/2016, 12:57:25 PM, LKSLIB\L1A\L2\FB1(IN)\CR12, 0de92d541c6b8e56be32a878563a666aa200e1d63d53f94f5ef3330128f5b310, UNDERGRADUATE STUDENTS, Lee Kong Chian School of Business, Bachelor of Business Mgmt, AY_2013, 0"

Frequency of count: 100,704

Name of Resource being looked up: SMU Library Entrance Turnstile, Internal Database

Parameters	**Description**	**Example**
Date	This is the date which the log is recorded when the user taps his/her matriculation card.	2/1/2016
Time	This is the timing which the log is recorded when the user taps his/her matriculation card.	12:57:25 PM
Device Name	This is the specific device being tapped on by the user.	LKSLIB\L1A\L2\FB1(IN)\ CR12
Email (hashed)	This is provided so that we can match the unique email to the corresponding ones in the proxy data logs.	0de92d541c6b8e56be32 a878563a666aa200e1d6 3d53f94f5ef3330128f5b3 10
User Group	This indicates which level of education the user is in, typically Masters or Bachelors programme.	UNDERGRADUATE STUDENTS
Statistical Category 1	This indicates the school that the user is from.	Lee Kong Chian School of Business
Statistical Category 2	This indicates the specific programme the user is undertaking.	Bachelor of Business Mgmt,AY_2013
Statistical Category 3	This indicates the year which the user is admitted into SMU.	AY_2013
Statistical Category 4	This indicates the year which the user is graduated from SMU.	0

WORK METHODOLOGY

Tools

Name of Tool	Version	Method of obtaining	Main uses of the tool	Open-Source?	What we could possibly use it for
Tableau	10.1	Downloaded from main webpage at www.tableau.com	Instantaneous visualization of data on dashboards via a drag-and-drop interface	No, attained through a paid Key by SMU during our Analytics Foundation module	Visualization of interesting insights from user behaviour
JMP Pro 13	13	Downloaded from SMU files directory provided by Prof. Kam	Statistical discovery software which enables interactive data visualization and analysis, mainly used for business analytics	No, attained through SMU for Analytics Practicum module	Text analysis on log files to develop interesting insights from user behaviour

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 the following (with thought process written below them)

Possible additional Data	Reasons	How may it impact our end results
Dates of public holidays	Amount of search queries may dip when users are at holidays	We can then explain the dip in search queries by users on certain dates in the months and account for them
Period of semesters (1, 2, 3A, 3B)	Amount of search queries may fluctuate depending on which semester it is	We can then observe how the amount of search queries fluctuates according to the length of the semester (Eg. Term 3A is lesser in duration than a regular Term, thus queries may be packed in a shorter period as opposed to a longer period in a regular Term where students have the luxury of time)
Periods in semesters where researches are mostly done	Amount of search queries may fluctuate when it is near project submission dates (Eg. 2 weeks before project submission	This may account for the sudden surge in different weeks of the regular Term

	seems to be the best time to get research done)	
Period of recess week	Amount of search queries may fluctuate during recess week	This may account for the sudden dip in recess weeks
User experience and User's general behaviour through a set of qualitative and quantitative questionnaire to add onto the 1-dimensional analysis from the log data	Users may use research for different purposes (eg. Law students may simply search for cases as specified in the course outline, while Business students may search for a wide range of topics to explore deeply as per their project requirements)	This can then make the otherwise 1-dimensional analysis from log data come to live as profiles of users whose behaviours are being analysed from their search queries, will be supported/complemented with the results from the qualitative/quantitative questionnaires

Data Preparation

As the Initial Dataset file is large (2.5Gb) and in .txt format, we will first find a software that can 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)

Types of analysis	**Description**
Association analysis	Whether there is an association between certain search queries that users search for
Visualization of association	Visualization that helps us to view the said association better
User behaviour analysis	Analysis of user behaviour to find possible correlations between groups of users
User clustering	Group users based on their behaviour

Risks and Limitations

Risks and Limitations	**Possible ways of mitigating**	**Back-up Plan**	**Testing Strategy**	**Presentation**
Analysis of log files not done by sponsors yet, thus we need to start on our own	Reading up of literacy research on how log files analysis is typically done	Seek expert help from Professors specialized in analytics, specially how to extract information from logs	Test on a small dataset of log files first, from a selected few database, then extend to include the whole database	Present findings on log files and how we extract information from them, and then include the whole database if sponsors/supervisors would like to view it
Missing search queries due to the design of the library logging system	Finding out where are the missing search queries typically from, which database it	Recognize the pattern for such missing queries, if not possible then present	Test on a small database to see whether missing search	Present findings on where the missing search queries typically are and how do we solve such a problem.

	occurs	findings as a critical limitation to the system and how to solve it by reading up on how such issues are typically solved and then recommend it to the sponsors	queries can be detected before extending it to include the whole database	

KEY STAKEHOLDERS

Project Supervisor

Prakash Chandra Sukhwal, Instructor (Analytics Practicum & Analytics Foundations), Masters of IT in Business (Analytics), SIS, SMU

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 (Please take note that this list may be altered per the progress of the project)
    a) Methodology explanations to handling the given dataset
    b) Visualizations to portray event sequence for each unique user
    c) 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

S/N	Task	Week1		Week2		Week3		Week4		Week5		Week6		Week7		Week8		Week9		Week10		Week11		Week12		Week13	
	Project Initiation																										
1	Source for data sponsor	ALL	ALL																								
2	Understand business domain	ALL	ALL																								
3	Acquire data	ALL	ALL																								
4	Create proposal	ALL	ALL																								
5	Update wiki	ALL	ALL																								
	Data Preparation																										
6	Explore data	ALL	ALL																								
7	Find out patterns			B&JX																							
8	Develop algorithms			B&JX																							
9	Clean data			B&JX																							
	Pre-midterm Analysis																										
11	Association analysis							JX&JR		B&JR																	
12	Visualization of association											JR															
	Midterm																										
13	Prepare Mid-term report													B&JX													
14	Update wiki													B&JR													
	Project Milestone: Mid-term Presentation(19th Feb)														ALL	ALL											
	Post-midterm Analysis																										
15	User behaviour analysis															ALL											
16	User clustering																	ALL									
17	Visualisation of users clustering																			JX&JR							
	Final Presentation Preparation																										
18	Consolidate findings																					ALL					
19	Prepare final report																					ALL					
20	Finalise wiki pages																							ALL			
	Project Milestone: Final Presentation()																									ALL	ALL
			Plan																								
			Actual																								
			Project Milestone																								

LITERATURE RESEARCH

Text Mining

Text Mining is “the automatic processing of natural language text data available in reasonably large quantities in the form of computer files, with the aim of extracting and structuring their contents and themes, for the purposes of rapid (non-literary) analysis, the discovery of hidden data, or automatic decision making” (Tuffery, 2011)[3].

Possible Methodologies and Best Practices

A useful step in progressing from a set of data to implementation of a text mining service over that data set is the use of paper prototyping methods (Tonkin et al., 2016)[4]. There are various approaches to design and prototyping available to text mining. Some of which are namely:

- Worked Examples, which help to express required functionality.
- Low fidelity mock-ups of interfaces and demonstrating workflows to users.

Working on a text mining project involves concerns on behaving responsibly on the web. Publishing the results of a content mining project requires decisions regarding the availability, presentation and format of the mined data (Tonkin et al., 2016)[4].

There are many ways to exploit the associations of semantic descriptions with text spans in a document collection, for the benefit of the user of a search engine (Tonkin et al., 2016)[4]. Text mining can help to create a valuable linked data resource if the semantic annotations go beyond the classification of text spans, and associate them with database identifiers (Hoffmann, 2007; Vanteru et al., 2008; McEntyre et al., 2011)[5,6,7]. The semantic annotations could also be searchable themselves.

From the user’s perspective, a full-text search often produces a very large result set, and various techniques exist to show the user ways in which the result set can be broken down into smaller, manageable sets. Faceting shows the distribution of a result set among subsets sharing metadata attributes and values, and is now a very popular tool in e-commerce (Tonkin et al., 2016)[4]. In addition, clustering of query search results can be derived on the basis of unsupervised algorithms that assign indicative labels to sets of lexically similar documents, but is only practical for a subset of the largest result sets (Tonkin et al., 2016)[4].

As the main focus for the document pre-processing phase, Natural Language Processing (NLP) techniques like statistical and machine learning approaches are required. Spiros Sirmakessis (2004)[8] has reported the sequential approach to the pre-processing of documents:

1. **Data Selection and Filtering:** First step to reduction of the large amount of data available, which helps avoid the overload related to the computationally intensive pre-processing and mining processes.

2. **Data Cleaning:** The removal of noise from the textual data in order to improve its quality.

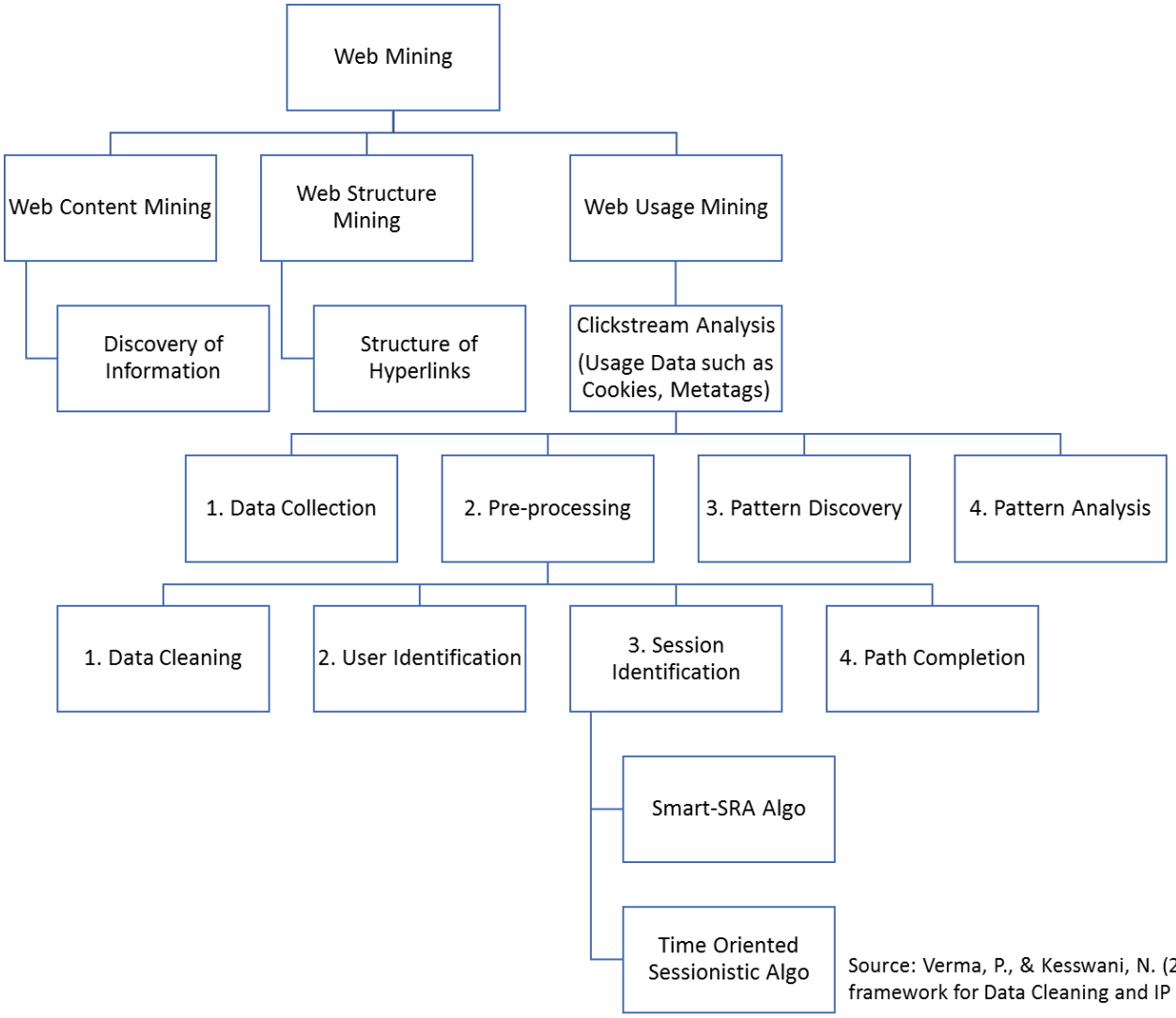
3. **Document Representation:** Dimensions of the full-scale feature vectors are associated with the words extracted out of the document (collection vocabulary).

4. **Morphological Normalization and Parsing:** NLP tasks such as stemming, lemmatization and Part-of-Speech tagging which aims at the production of canonical surface forms.

This is supported by the web cleaning model by Verma, P., & Kesswani, N. (2014)[9] whereby the main steps of the model coincides with the sequential approach to pre-processing of documents. (The model is shown on the next page)

Spiros Sirmakessis (2004)[5] also suggested that text mining essential belongs to descriptive data mining, however, predictive techniques such as trajectory identification and trend analysis are also investigated. Following are the main data mining techniques for textual data:

- Clustering Techniques
- Classification
- Relation Extraction: Relations between individual features in the vectors, which are substantial important information required for mining.
- Entity Extraction: Assigning some pre-defined labels to the textual entities that hold a given interesting semantic property.



Source: Verma, P., & Kesswani, N. (2014). Web Usage mining framework for Data Cleaning and IP address Identification.

REFERENCES

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

2) 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.

3) Tuffery, S., 2011. *Data mining and statistics for decision making.* Chichester, West Sussex: Wiley.

4) Tonkin, E. L., & Tourte, G.J.L., 2016. *Working with Text: Tools, Techniques and Approaches for Text Mining.* Cambridge: Elsevier Ltd.

5) Hoffmann, R., 2007. *Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds.* Curr. Protoc. Bioinformat. http://dx.doi.org/10.1002/0471250953.bi0116s20. Chapter 1, Unit 1.16.

6) Vanteru, B.C., Shaik, J.S., Yeasin, M., 2008. *Semantically linking and browsing PubMed abstracts with gene ontology.* BMC Genomics 9 (Suppl. 1), S10. http://dx.doi.org/10.1186/1471-2164-9-S1-S10.

7) McEntyre, J.R., Ananiadou, S., Andrews, S., Black, W.J., Boulderstone, R., Buttery, P., et al., 2011. *UKPMC: A full text article resource for the life sciences.* Nucl. Acids Res. 39 (Database issue), D58-D65. http://dx.doi.org/10.1093/nar/gkq1063.

8) Spiros, S., 2004. *Text mining and its Applications: Results of the NEMIS Launch Conference.* New York: Springer-Verlag Berlin Heidelberg.

9) Verma, P., & Kesswani, N. (2014). Web Usage mining framework for Data Cleaning and IP address Identification.