SMU
SINGAPORE MANAGEMENT
UNIVERSITY

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

<u>Proxy Log Data</u>

```
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=Sea
rchVolumeSOLR%3Finput%3D%2528The%2520Great%2520Peace%2529%26div%3D7%2
6f_size%3D600%26num_results%3D10%26handle%3Dhein.journals%252Ffchlj23%26colle
ction%3Djournals%26set_as_cursor%3D19%26men_tab%3Dsrchresults%26terms%3D%25
28The%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
    a) Methodology explanations to handling the given dataset
    b) Visualizations to portray event sequence for each unique users
    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	🟨	🟩																								
2	Understand business domain	🟨	🟩																								
3	Acquire data	🟨	🟩																								
4	Create proposal	🟨	🟩																								
5	Update wiki	🟨	🟩																								
	Data Preparation																										
6	Explore data	🟨	🟩																								
7	Find out patterns			🟨																							
8	Develop algorithms			🟨																							
9	Clean data					🟨																					
	Pre-midterm Analysis																										
10	Descriptive analysis							🟨																			
11	Association analysis									🟨																	
12	Visulisation of association											🟨															
	Midterm																										
13	Prepare Mid-term report													🟨													
14	Update wiki													🟨													
	Project Milestone: Mid-term Presentation(19th Feb)														🟥	🟥											
	Post-midterm Analysis																										
15	User behaviour analysis															🟨											
16	User clustering																	🟨									
17	Visualisation of users clustering																			🟨							
	Final Presentation Preparation																										
18	Consolidate findings																					🟨					
19	Prepare final report																					🟨					
20	Finalise wiki pages																							🟨			
	Project Milestone: Final Presentation()																									🟥	🟥

🟨	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).

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). 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 (Tourte et al., 2016).

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). 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). 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 (Tourte et al., 2016). 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).

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) 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.

Spiros Sirmakessis (2004) 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.

REFERENCES

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.

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.

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.

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

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

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

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.