Li Ka Shing Library - Mining for Insights from Users' Database Request Log

Dina Heng Li Gwek | Lu Ning | Song Rui

Sponsor: LKS Library Analytics Team

Supervisor: Kam Tin Seong

TABLE OF CONTENTS

OVERVIEW

Summary

This project aims to provide insights on users and ebook databases for Li Ka Shing Library Analytics Team.

Sponsor Background

Li Ka Shing Library is the center for academic and professional knowledge resources and services that support the research and learning needs of the SMU community. With connection to over 180 electronic databases and over 400,000 printed and electronic books and journals, it aims to provide seamless access to information using innovative and leading edge technology.

Project background and motivation

The library subscribes to eBook platforms with contents from an array of publishers. These databases provide contents that have largely enriched the library's resources and make an integral part of the library repository.

When a student user requests contents from the databases, the request goes through the library's proxy server. The proxy server captures a digital trace for each user request, which contains request url, user ID, and user agent. With the aim of providing easy services to users, the management hope to better understand the usage patterns of the databases. The challenge is to programmatically extract the meaningful user inputs within billions of request record, as most records are irrelevant to the project objectives.

Our main focus of the project is to understand the usefulness of the library eBook databases in fulfilling the student queries. By analysing the proxy trace, we can define the characteristics of the users as well as examining the usage rate of the database. Since the dataset is not static, we aim to provide a processing pipeline to help the sponsor in looking for new findings with new enrolled students in the future.

The reasons that we chose this project are many folds. The expected results are a draw to this project. Analysing "web usage logs offer a direct and immediate record of what people have done on a Web site—not what they say they might do but what they have actually done"(Black, 2009). As SMU students, we are interested in knowing the analytics results as it is easy to relate to ourselves. Besides, this project would provide a learning opportunity for us to put many data analytics skills into practice. With a flexibility to decide how far we want go, this project is difficult enough to make it a fruitful learning experience, but not too much to burn out ourselves. The challenges are largely on data wrangling, which requires the team to extract relevant data that reflect user behaviour from web logs, and to transform the data into a proper format. To do this, considerable amount of programming with R or Python would be necessary. Working on this project, the team will also be able to learn the domain knowledge in library data analysis

and library management. All these skills will equip us with advanced data analytics skills and prepare us to for a career as data scientists.

Project objectives

1. Understand the characteristics of database and users requests.
a. Help the management understand and take actions on each database by profiling e-book databases. The considered parameters include
    i. user profiles (faculty, program, year)
    ii. number of requests
    iii. popular requested items.

    The profiling will be done on multiple timescales (e.g. by hours in the day, by days of the week, or by weeks) to identify chronological patterns. The analysis result will help decision makers understand *who* requested for *which items* from each database. By slicing data on multiple timescales, the team will be able to identify the peak periods and general trend of requests for each database.

    b. Help the management better understand the user behaviours by profiling users; identify special e-book usage patterns for students from different faculties or with different academic performance. The usage patterns are chronological patterns, device used, and sites requested. This would help the management to better understand the users, and devise better approach to improve service for each type of users.

2. Timing of student access the database is another aspect of our focus, which is to dive deeper into the finding and patterns regarding the databases users. A focus we will be taking is to compare the behaviour pattern of dean's-list students against other students. When do they access the materials? One possible aspect is to find out any group that is particularly in favour in last minute work.
3. Consistency of the student access the resources is a continuous aspect of the previous objective as it is useful to know how students need the materials across the entire semester. Whether it is widespread or intensely concentrated on a certain period. So, it can give insights to manager to understand the student behaviour pattern and give suggestion on workshop planning and resource preparation.
4. Within the sessions of all the users, we can carry out ‘Market Basket Analysis’ to sniff out the popular combination of e-books being viewed and downloaded. The purpose of such action is to potentially create a foundation for e-resources recommendation in the future.

DATA FOR ANALYSIS

The data we will work with is request log data (i.e. digital trace) and student data. Request log is a NCSA Common Log Format data with billions of records captured by the library's URL rewriting proxy server. This data set captures all user request to external databases. The data includes dimensions of user ID, request time, http request line, response time, and user agent. The student data, specifying faculty, admission year, graduation year, and degree program, is also provided for the team. For non-disclosure reason, the user identifier - emails - are obfuscated by hashing to a 64-digit long hexadecimal numbers. The hashed ID will be used to link up two tables. Please refer to appendix for the complete data dimensions and samples.

There are users other than students (e.g. alumni, staff, visiting students and anonymous users), but the scope of this project is only limited to students because of the availability and volume of student data.

METHODOLOGY

The team was suggested an open-source proxy log analysing tool ezPAARSE. ezPAARSE can extract request time, user identifier, digital object identifier (DOI) and resource type from the request log. However, just using this tool alone is not enough, as it is not able to extract the content title or user query from the url. Based on our own expertise and the features of the tools, we decided to use R and Python for data processing and reporting.

WORK SCOPE

To achieve the project objectives, the team will go through the typical data analytics steps. The following steps serves as guideline for the project roll-out:

1. Data exploration: understanding of data
2. Data preparation: cleaning, reformatting of data
3. Data analysis: using multiple data analytic methods to generate insights
4. Reporting: giving actionable recommendations drawn from the summarized information and insights
5. Data processing pipeline: an automated data processing pipeline that produces new insights from new data set
6. Post solution analysis: how one should further improve on this study

At the end of the project, the findings will be presented formally to the library analytics team. The pipeline will also be delivered, with step-by-step instructions.

Data exploration

After objective establishment, the problem is analysed by studying the available data. The purpose of this step is to 1) uncover the underlying structure, 2) extract relevant variables, 3)

test assumptions. When doing data exploration, the team will be able to determine the feasibility of the objectives, and make necessary changes with the sponsor.

As our team have already started exploratory analysis on a given sample data of one month's requests, which contains 6.62 million records, we have already identified some features as follow:

1. Over 25 percent of the requests are for web resources (e.g. js, gif).
2. 87 percent of the requests are *HTTP GET* request
3. Request are unevenly directed to the databases. (stdev = 46.6, avg = 10.2)
4. Multiple encoding of search phrase in request url, based on the database
5. Requested items are usually serialised in their own way
6. 10 percent of the requests point to internal URL
7. 16 percent of the requests have a status code other than 200

In short, the only a small proportion of web log reflect user request. Literature review has also shown that hits are not very informative because they can vary widely because of site's graphical design and architecture instead of the actual usage. (Jana & Chatterjee, 2004) This implies that to obtain insights, we have to infer users' actions from the log instead of a simple statistic on the log itself.

Data Preparation:

The given data is all well-formatted, but we have identified necessary preparation of data is required in following areas:

1. Remove irrelevant requests such as those to web-resources or general web pages
2. Student and request data is collected from different sources and stored in separate tables; a joining of the data is required before student request patterns can be analysed.
3. Extract date time components from CLF date time string to slice data
4. Extract domain for each request to identify the database requested
5. Split the data into three groups: 1) with explicit search phrase in URL 2) without explicit search phrase but with other forms of reference to a title
6. Extract the requested title/search phrase
7. Infer the category of resource from the requested title/search phrase

Data analysis

1. Slice records by various time frames, and user attributes, and count the number of records. By analysing the number of requests under each dimension, we will be able to capture the characteristics of the databases.

2. We are able to text mine the popular topics of all the search result via categorical analysis, as normal clustering analysis required sequential data, but the e-book's' title and tags are discrete data, hence we need to do categorisation before conduct pattern findings on them.

3. K-means clustering to profile students on dimensions of

a. user agent
b. most frequently used database
c. most frequently requested resource category.
d. faculty
e. if the user is in Dean's List

This result will help the management better understand the user behaviours for different segments of students.

4. Perform Market Basket Analysis on the titles viewed by a student in a session. The analysis results can be potentially used for recommendation system.

Reporting and data processing pipeline

At the end of the semester, the analysis results will be presented with interactive charts. A formal report will be delivered to the library analytics team. The report will consist of all useful findings from our study, and actionable recommendations to the library. The team will advise on areas of interest for further studies. A data processing pipeline consisting of the necessary scripts and programs will also be delivered, with instructions on setup and execution of the process. The library analytics team will be able to rerun the analysis on new request log data.

GAP ANALYSIS

There are a few areas the sponsor wants to explore but is limited by availability of relevant data. For instance, it would be helpful to find out the cost on database subscription regarding the number of user requests. However, the subscription schemes are rather complicated and unfit for analysis. Another example is to establish association between a student's modules and e-book requests. This would shed light on the needed e-book resources for a given module. However, this association is not obtainable due to unavailable student modules data.

Besides, to achieve the objectives, an inevitable step is to extract the search phrases or requested titles from request URLs. As each database site has different URL pattern for GET requests, and for some databases the requested titles are represented by their IDs, accurately extracting the title from URL is difficult, especially the request result is not reproducible due to access right constraint. Missed or inaccurate extractions may lead to very biased analysis result.

PROJECT TIMELINE

The team have proposed weekly meetings with Supervisor and monthly meetings with Sponsor (last Wed of every month. Tentatively set for Week 4/8/12/16).

On top of these meetings with our sponsor and instructor, we also have a weekly meeting minutes which will be uploaded on Wiki every Sunday night. On top of that we will upload the wiki every week regarding the project's progress as well.

The table below shows the planned schedule.

Tasks	W0	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
Milestone Submission: 1 Jan (Proposal Deadline)																	
Data Gathering and Scoping																	
Gather data	D, L, S																
Finalise Requirements with client	D, L, S																
Scope project	D, L, S																
Research and Preparation																	
Explore software	D, L, S																
Finalise Proposal	D, L, S																
Create and Update Wiki Page	D, L, S																
Milestone Submission: 19th Feb (Interim Report) + Interim Presentation																	
Data Cleaning																	
Data collection		D, L, S															
Data cleaning and restructuring		D, L, S															
Resolve/remove incomplete data		D, L, S															
Data Modelling																	
Stage 1: Exploratory Analysis				D, L, S													
Stage 2: Clustering				D, L, S													
Stage 3: Time Series/Seasonality				D	D	D	D										
Interim Preparation																	
Gather feedback from client					D, L, S				D, L, S								
Prepare Interim Report and Slides							D	D									
Update Wiki		S	L	D	S	L	D	S									
Milestone Submission: 2nd April (Abstract Paper Submission), 20th April (Full Paper Submission)																	
Application Building																	
Code the application									D, L, S								
Testing of application												D, L, S					
Gather feedback from client													D, L, S				D, L, S
Iteration																	
Adjust analysis													D				
Refine results to improve clarity													L, S				
Final Preparation																	
Prepare Research Paper															D, L, S		
Prepare Poster																L	
Update Wiki									L	D	S	L	D	S	L	D	S
Conference Day: 22nd - 23rd April																	

EXPECTED LEARNING OUTCOMES

We would like to acquire the knowledge on data processing skills with R data analysis programming language, as well as understand the how to approach web log data in massive quantity. We believe it is useful for us in the long run as not only we can utilise the R programming skills in many other data analysis task but also process proxy log data in other industries such as ecommerce and social media platform.

Furthermore, by keeping the current standard of our documentation practice, it will also reinforce our project management skills. It helps us to better evaluate our true ability and optimise our working performance by preventing us from over or under promising.

REFERENCES

1. Black, E. (2009). Web Analytics: A Picture of the Academic Library Web Site User. Journal of Web Librarianship, *3*(1), 3-14.
2. Jana, Sanghamitra, and Supratim Chatterjee. 2004. Quantifying Web-site visits using Web statistics: An extended cybermetrics study. Online Information Review 28(3): 191–99.

APPENDIX1 - DATA TABLES

USER		
Metadata	**Example**	**Description**
Email (email)	feb0e4d05b236c0bcc0c7331dc754921cf9189c4c1317b0b112696fcf68cd2f8	64-char hash
User Group(user_group)	UNDERGRADUATE	Many “others”
Category 1(school)	Lee Kong Chian School of Business	Full name of school
Category 2(degree_program)	BSc (Economics)	
Category 3(admin_year)	AY_2011	
Category 4(grad_year)	GY_2012	Program completion year

REQUEST		
Metadata supposed	**Example**	**Description**
ip	59.189.71.33	Might be used to locate requester
Session id	tDU1zb0CaV2B8qZ	Define the online duration
email	65ff93f70ca7ceaabcca62de3882ed1633bcd14ecdebbe95f9bd826bd68609ba	
time	[01/Jan/2016:00:01:33 +0800]	

request_line	"GET http://heinonline.org:80/HOL/LuceneSearch?terms=The+Great+Peace&collection=all&searchtype=advanced&typea=text&tabfrom=&submit=Go&all=true HTTP/1.1"	method uri httpVersion
status_code	200	
response_size	2327	bytes
user_agent	"Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36"	