SIT792 Minor Thesis

Customer Behavior Analysis using Weblog data

Minor Thesis

Mohitkumar Rangholiya
Student ID: 215410048
Supervised By
Dr Gang Li

October 3, 2017

DEAKIN UNIVERSITY

List of Figures

List of Tables

Contents

Acknowledgment

I would like give a sincere thanks to my supervisor, Prof. Gang Li, for the patient guidance, encouragement and advice he has provided throughout my time as his student. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. Not only that, but his continuous inspiration towards research work gave me immense interest as well as future research insights.

I must express my gratitude to Hiran and Adekola Oluwatade, my colleagues, for her continued support and encouragement. I was always helped by him in solving technical errors during implementation work, and by the patience of my brother who experienced all of the ups and downs of my research.

Completing this work would have been all the more difficult were it not for the support and friendship provided by the other members of the School of School of Information Technology, Deakin University. I am indebted to them for their help.

Abstract

In this era of Internet, World Wide Web is increasing its size day by day. Most of the industries are using this technology directly or indirectly. Hospitality and tourism sector is one of the interesting sector of study. Literature synthesis proposed here focuses on identifying behavior of online visitors of a hotel website using Web Server logs. To achieve this, various data analysis techniques including data cleaning, supervised learning, Frequent Pattern Mining and contrast analysis will be used to analyze the visitors visiting path. Moreover for training and testing, data have been taken from the Server Logs of a hotel website. This research of insights will help Hotel to improve customer satisfaction using their online portal and hence this will help the Hotel or any Hospitality Industry to increase their online customer base simply by utilizing their web server logs.

The goal of discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the website users. This leads to the most interesting to least interesting web-page of the website. Moreover this information can be used for website improvements, offers and advertisements, marketing campaign and other business applications. This thesis presents preprocessing of web log data by identifying user sessions and mining frequent patterns to identify website user's behavior. Various powerful insights were captured from the analysis and used to explain contrast between web access patterns by different sets of users.

Chapter 1

Introduction

This research thesis focuses on applying data analysis techniques in tourism sector using web server logs. As part of research thesis, this literature synthesis report is the review of the related researches carried out in this area. To understand or research of any significant topic, it is the basic requirement of studying all the related past researches and then start working on challenges or issues acquired. In this report, we will first create the background of the research. Then after we will study key research areas and continuously, we will review all the related researches leading to our aim. Once we will be familiar with our research scenario, our main focus will be on methods and methodologies we will be using to carry out our research. After having discussion of this we can figure out key challenges we come across and future research questions to be addressed. Last but not the least, we will jump on exciting part of any report which is conclusion.

With the tremendous growth of World Wide Web, each and every sector use websites for branding and selling their services. The fast growth of information technology in common and the Internet in particular has enormously upgraded the tourism and hospitality sector [9]. The modern Tourism and Hospitality enterprises has highly adopted e-commerce to accomplish their business goals and so, the process of building, maintaining high quality and use-friendly websites has become crucial to strengthen customer relationships and to excel their profits [8]. Number of researches have been adopted the study of accessing the website effectiveness and design. For Web page design, both the functionality and usability of Web pages are thoroughly considered [12].

This study will be helpful to web-designers for designing effective and customer oriented web-pages. Effective website design means that the website is free of barriers to online transactions like anonymity, lack of communication and electronic payment options [20]. The effectiveness of hotel websites can also be measured from the perspectives of easy to use and right content [22], which can be reflected into *functionality*and *usability*. Here website functionality, refers to the degree of information provision about the websites services/products, while usability refers to what extent a website is efficient and enjoyable for its products/services being promoted [22]. Another study in this area introduced a tool that evaluates destination websites according to four quality factors integrated from

existing tools: *information completeness*, *credibility*, *usability*, and *persuasiveness* [17]. Other studies have carried out an investigation of factors affecting customer selection of online hotel booking websites [11].

Website visitors leave their access behavior while visiting websites. These websites are hosted on web servers. Web Server stores user's web access information in Log files and thus Web log data plays significant role to understand web surfer's access behavior. It is widely accepted fact that When you know someone better, you can serve them better. Similarly, Online website visitors are hotel's prospective customers and if hotel management is aware of services that are most interesting to it's customers then it can definitely help them to improve their customer's online experience as well as their overall satisfaction about the hotel. Prior studies of web log mining has achieved certain degree of success in the direction of identifying visitor's access behavior. The number of previous attempts to understand web server logs, unfortunately, is very limited in the existing tourism and hospitality literature [10]. But with the help of modern data mining techniques, this paper aims to improve the use of web log data in Tourism and Hospitality sector to identify online visitor's behavior.

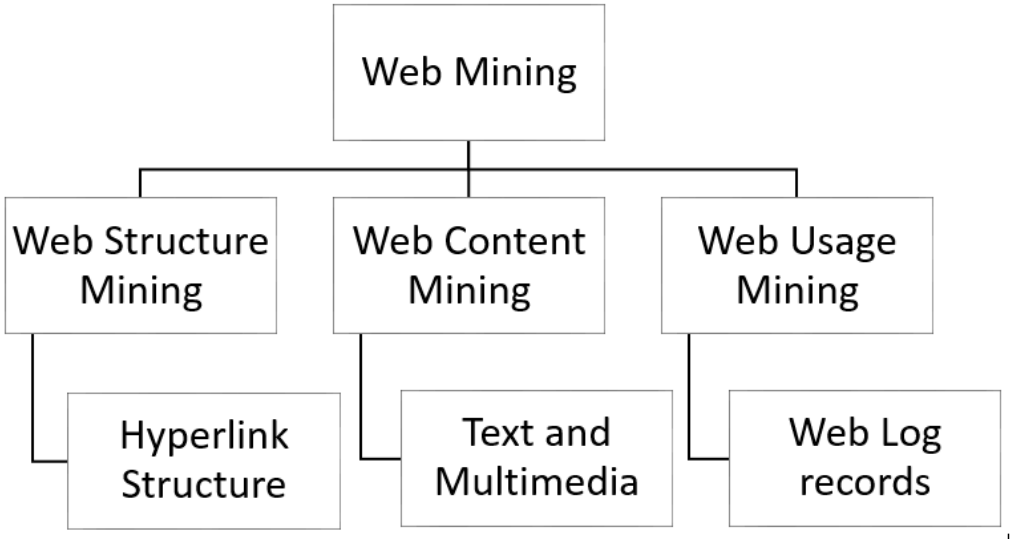

Figure 1.1: Web Mining Techniques

Web mining is the application of data mining for knowledge discovery from the Web. As described in Fig.1.1 Web mining can be classified into three various categories [7]. Web Content Mining is the process of capturing useful information from Web documents. This information may consists of text, images, audio, video or structured records like tables and lists [18]. Web Structure Mining deals with the internal linked structure of the website. The major goal of Web Structure Mining is to shape structural abstract of the website [18]. While Web Usage Mining is an application of data mining techniques to discover useful insights. Identifying hidden patterns from Web log data is an useful application of Web usage mining.

Chapter 2

Background and Motivation

To begin with, one should understand what is this research about and why it is proposed here? Any research can be justified using evidences provided. In this scenario, our research is based on practical implementation using various technologies in data analytics. As described before, the log dataset is taken from web server of hotel website. The research aims to identify online visitor's behavior in terms of how frequently users visit the website, various patterns of visiting website, visitors from deferent demographic areas and their behavior, Finding out the best server maintenance time, finding out ideal and busy hours and days for Web Server and similar useful insights that can be taken into consideration for improving customer satisfaction and overall website performance.

2.1 Motivation

Hotel Y wants to start use of the server log data for their internal decision making process. Hotel Y has got continuous flow of log data from it's web server which means the volume of data is being increased everyday and hence it is important to store and utilize this data in an efficient way. It's always challenging perform analysis on large volume of data. So having thought of utilizing the data, can be transformed to some useful insights from the data which can be taken into consideration while decision making process. The main motivation behind this is to make use of server log data to gain actionable knowledge.

2.2 Company Background

The analysis undertaken here is being performed on the dataset provided by Hotel Y. The Hotel Y is well-known for it's beautiful view, fabulous food services and their customer satisfaction. Hotel Y offers variety of dining facilities, various types of rooms, facilities like business meetings and events, Wedding and occasion venues. The goal is to provide customers with an Internet experience that delivers the information, resources and services that are most relevant to them. The management of Hotel Y is interested in

exploring it's business in international tourism. It has plenty of worldwide visitors who directly use website to book various facilities provided by Hotel Y. The server of Hotel Y website generates logs for each online interactions by users. So the management of Hotel Y aimed to use these log-files for their decision making process.

Chapter 3

Research Problem Analysis

Before initiating any research work, it is compulsory to identify research challenges and Research Questions that are to be answered through the research. Research Challenges allow us to brief the scope of the research by using certain questions. The below section describes the Challenges and Questions identified to do Customer Behavior Analysis from the Hotel Web log data.

3.1 Research Challenges

Storage and access of big log data: Storing and accessing Bigdata is an emergent research area. Web server of any busy website generates Big Log data and when it comes to analysis, it's first thing to be taken care. This problem must be answered properly to start the actual research problem.

Exploratory Data Analysis using Python: This is a simple but very useful part of our research. Data Exploration is easy in python but as we mentioned before, Accessing *Bigdata* using *Python* is again a challenge in our research.

Identifying Frequent Web page Pattern: The most important analysis of our research is identifying frequent pattern of user's web visits. Log data contains whole *uri-stem* which contains webpage name, finding subsequent web-pages in order seems difficult task and should be focused more.

Contrast Analysis on logdata: The term Contrast Analysis here is comparing visitor's behavior from two or more groups. These groups may taken as visitors from different countries or various states of a particular country.

3.2 Research Questions

As this research aims to many important analysis, they can be categorized in various research questions and then using data analysis techniques, they can be answered further.

1. How and Where to store the log data for analysis?
2. What tools and technologies are efficient for Big log data?
3. How to identify unique users and sessions in the log data?
4. What can be the useful business insights for Online Hotel Websites?
5. How can frequent web-page visits be identified from web-log data?
6. How can Sequential Frequent Pattern Mining be performed on log data?
7. How contrast analysis be useful for finding out online hotel web service purchasers?

Chapter 4

Key Related research review

In this modern era of Internet, World Wide Web is growing at incredible rate. It has become customary today, It allows users to interact and collaborate with others in very easy and satisfactory way. The different types of data have to be managed appropriately and efficiently so that, it can be used in various decision making in the real world problems. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers [5]. Web Mining is one of the popular type of data mining methods to extract useful information based on users' needs, under web mining, web usage mining is one of the application of data mining technology to extract information from web-log to analyze user's behavior on visiting websites [14]. Web mining can be categorized into three major types, which includes Web content mining, Web structure mining and Web usage mining [7].

Web content mining refers to the finding useful information from online resources [1]. Web structure mining can be defined as the process of finding the structure hyper-links within the Web [5] [6]. Web usage mining refers to the task of finding useful activities or navigational patterns of the users while they are browsing the Web. The significance of understanding navigation of the user is to improve websites for better user-experience or server performance. So the major sources of data we get here is server logs. There are major three categories of these log files. Log files stored at server side, client side and proxy servers side [5].

Todays popular websites are visited by thousands of Web users each day. Visitors leave behind their visiting behavior in the form of Web logs, which is monitored and saved by web servers. By analyzing these access Web logs recorded at these servers as well as proxy servers, it is possible to know the behavior of Website users. Moreover the interesting fact is that, It can be further possible to use the learned knowledge to serve the users better [21]. Specifically, out of this learned knowledge, useful kind of knowledge which can be applied immediately, is called actionable knowledge [21].

4.1 Web Server Log Files

Web log files are files which are generated by the web server and it contains all the information about the users browsing activities on the web site [16]. Web log files are created automatically on every single user click on website and it is in text format with extension of '.log' format, most of the times and the size varies from 1KB to 100MB [14]. Web servers collects large volume of log information in log files. These logs contains user's IP address, date, time, user agent, request type, server status, server sub-status client-status, client sub-status etc. These data can be combined together as a single text file, or further divided into various logs like access log, referrer log, or error log. However, user specific information is not stored in the server logs [19].

There are major three types of logs:

1. Web Server Logs
2. Proxy Server Logs
3. Browser Logs

4.1.1 Web Server Logs

History of web page access requests is kept up as a log document. Web servers are the exorbitant and the most widely recognized information source. They gather substantial volume of data in their log records. These logs contain name, IP, date, and time of the demand, the demand line precisely originated from the customer, and so on. These information can be bound together as a solitary text document, or, on the other hand partitioned into various logs, similar to access log, referrer log, or error log. [19]

4.1.2 Proxy Server Logs

It goes about as a mediating level of getting lies between user browser and web servers. Intermediary storing is utilized to diminish the stacking time of a page and also the lessen organize activity at the server and client side. The genuine HTTP access requests from multiple users to multiple web servers are followed by the intermediary proxy server. These proxy server logs are utilized as a data source for analyzing fake visitors behavior gathering of unapproved access requests sharing a common proxy server [19].

4.1.3 Browser Logs

On client side, web browsers are responsible for collecting activity logs using JavaScript or Java applets. To implement client side data collection, user cooperation is must in this case. Web server logs are used in the web page recommendation to improve the E-Commerce usability [15]

4.2 Web Log File Format

As in our analysis the major source of data is server logs, we should know about major three types of log file formats.

- W3C Extended log file format
- NCSA common log file format
- IIS(Internet Information Services) log file format

Out of these three, our focus is on understanding Microsoft IIS log file format. It is non-customizable ASCII format used to record more information than the NCSA common format but less than the W3C format [7]. A sample IIS log file format is shown in Fig. 4.1.

```
1 #Software: Microsoft Internet Information Services 7.5
2 #Version: 1.0
3 #Date: 2014-08-01 00:00:25
4 #Fields: date time s-ip cs-method cs-uri-stem cs-uri-query s-port
  cs-username c-ip cs(User-Agent) sc-status sc-substatus sc-win32-status
  time-taken
5 2014-08-01 00:00:25 10.130.0.12 GET /LocationContacts.aspx - 80 -
  202.140.108.99 Mozilla/5.0+(
  iPhone;+CPU+iPhone+OS+7_1_2+like+Mac+OS+X)+AppleWebKit/537.51.2+(
  KHTML,+like+Gecko)+Version/7.0+Mobile/11D257+Safari/9537.53 200 0 0 19
```

Figure 4.1: IIS Log file format

Chapter 5

Methods and methodology

To accomplish this analysis task, various tools and data science technologies have been used efficiently. A standard approach has been followed through out the process of implementation. In this chapter, dataset has been described and explained briefly. It is followed by Methodology by using the Fig. 5.1 which elaborates step by step process of pattern identification from web log data. Two major steps names User Identification and Session Identification have been explained with detailed algorithms.

This research is more focused on technical implementation of the analysis. Various tools have been used efficiently to carry out the task and also many technical resources [3] [4] [13] have been used to understand and formulate the research problem. Table 5.1 describes all the important tools used to carry out the research.

Table 5.1: Tools and Technologies

No	Name	Type	Application
1	Amazon S3	Cloud Storage	To store Web Log Data
2	Python 2.7	Programming Language	To write scripts
3	Apache Spark 2.1.0	Big Data Platform	Analysis on Big data
4	Databricks	Python Spark Execution Platform	To execute python scripts
5	Graphviz	Graph library	To generate graphs

5.1 Dataset Description

As described in previous section, the raw dataset is provided by Hotel Y and the data is originally generated by Web Server. The table 5.2 describes basic information about the data. Server continuously generates this logs from website interactions. But here for the analysis purpose, log-data that server has generated between August, 2014 to August,

2015 has been taken as raw dataset for this analysis and hence this report highlights all the analysis performed based on this 13 months data.

Table 5.2: Dataset Description

Properties	**Value**
DATA SET NAME	Hotel Y Weblog Dataset
DATA SIZE	15.8 GB
NO. OF ATTRIBUTES	14
NO. OF DATA RECORDS	73368256(394 .log files)
DATA SOURCE PROVIDER	Hotel Y
PERIOD	August,2014 - August,2015
DATA PRIVACY	Private

Fig. 5.1 depicts the flow of Web log analysis. As it can be clearly seen that, the input of this process is the log files. The data must be preprocessed to have the appropriate input for the data mining algorithms. The various data mining methods need different input formats, hence the preprocessing provides three types of outputs. The frequent patterns mining phase needs only the Web pages visited by a given user and so, sequences of the pages are irrelevant in this case. Also the revisits of the same pages are ignored, and the web-pages are ordered in a predefined order. In sequence mining, the original ordering of the web-pages is also significant, and if a particular page is visited more than once by a specific user in a user-defined time duration, then it is relevant as well. Hence the preprocessing step of this analysis provides the sequences of Web pages by users or user sessions.

5.2 User Identification

In this step, each unique user is identified using deferent IP addresses. The algorithm 1 describes step by step process of identifying users in web log data. This algorithm aims to identify users so this passes each log entry, check for the user IP and search in the list of unique IP. If it is not in the list then it adds that IP as a unique user and assign new User-ID in the list else it assigns the existing User-ID.

5.3 Session Identification

Session can be defined as a sequence of the web pages visited by a specific user during single visit. session identification takes place based on predefined sessions timeout by the server. Most commonly the session timeout is taken as one hour. Here the algorithm 2 explains step by step process of session identification. It takes User-ID table as input

Table 5.3: Attribute Description

Attribute Name	Data Type	Data Subtype	Description	Examples	Additional Notes
date	MC	DATE	The date on which the activity has occurred.	"2014-08-01"	YYYY-MM-DD
time	MC	DATE	The time at which the activity has occurred.	23:59:19	time is in coordinated universal (UTC)
s-ip	CN	ADDR	The IP address of the server on which the logs were generated.	10.130.0.12	It remains unique as this data has generated from only one server
cs-method	CN	STR	Requested action/method from the client	GET	GET and POST are most common http action methods
cs-uri-stem	CN	URL	URI stem information, which is the target of the action requested by the client	*/images/arrow_prev_cld.gif*	URI: Universal Resource Identifier
cs-uri-query	CN	STR	Query if any, that the client was trying to perform.	Yic=0	item=4
s-port	CN	STR	Server port number that is configured for the site.	80	80' is default server port so it's common value of s-port
cs-username	CN	STR	The name of the authenticated user that accessed the server.	-	Anonymous users are indicated by a hyphen
c-ip	CN	ADDR	IP address of the client that made the request.	80.179.11.217	IPV4 address
cs(User-Agent)	CN	STR	The browser type that the client used.	Mozilla/5.0+ (Windows+NT+6.1; +WOW64;+Trident/7.0;+rv:11.0) +like+Gecko	It comprises of the details like browser name,device name,browser version
sc-status	CN	STR	The HTTP status code. A value of 200 indicates that the request has been fulfilled successfully and 404 is for server error.	200 , 404	sc-status codes are mostly used to define the server side errors or success
s-substatus	CN	STR	The sub status error code.	0	
sc-win-status	CN	STR	The windows status code.	0	A value '0' indicates that the request was fulfilled successfully.
time-taken	MC	CURR	The length of time that request took to be respond by the server.	101	It's in milliseconds.

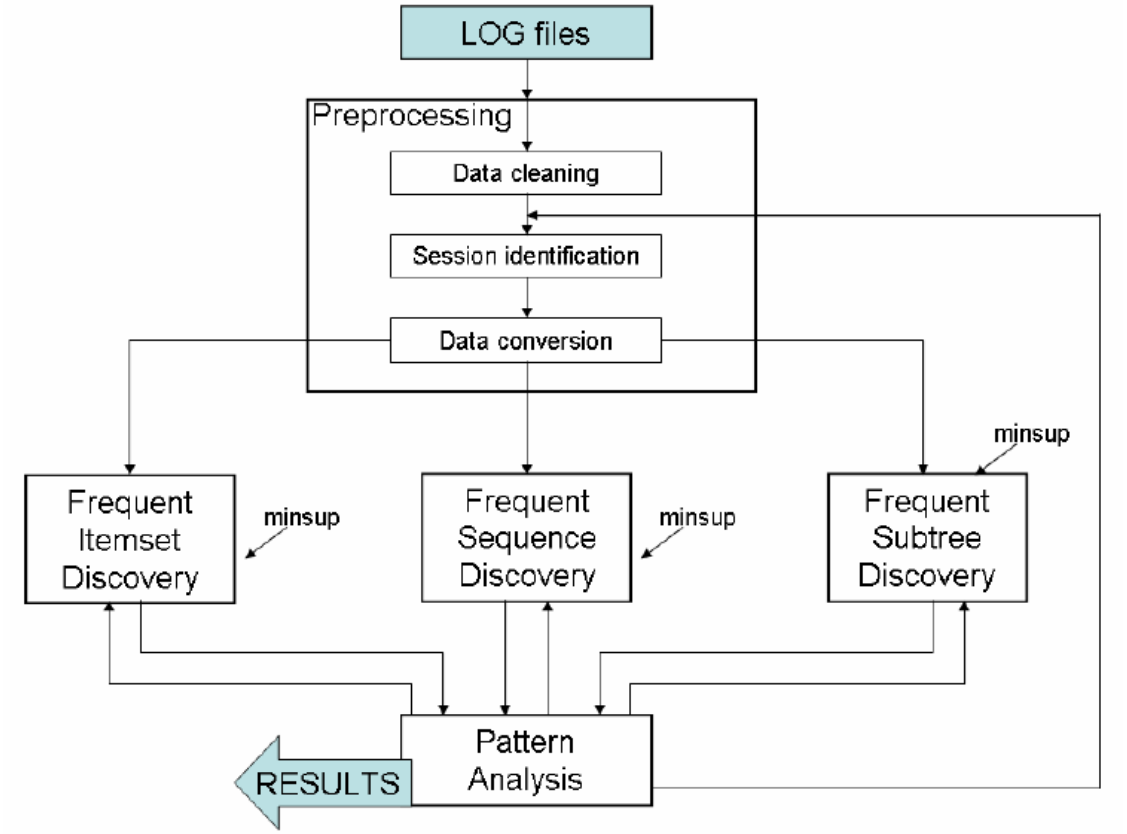

Figure 5.1: Process of Web Log Analysis [5]

Algorithm 1 User Identification in Weblog data [15]

Require: Refined Log Table
Ensure: Identified Users.
1: Read records in Log table
2: **for** each record in dataset **do**
3: **if** current IP is not in ListOfIP **then**
4: add the current IP in ListOfIP and mark whole record as a new user and assign User-ID
5: **else**
6: Assign the old User-ID.
7: **end if**
8: **end for**

and read through each records. records that has the time-stamp within an hour placed in one session and it assign new Session-ID. Similarly, if records found to be in the same duration as existing session for a specific User-ID, then it assign old Session-ID.

Algorithm 2 Session Identification in Weblog data [15]

Require: User identified table
Ensure: Identified sessions.
1: Read records in Log table
2: **for** each record in dataset **do**
3: **if** *time_required* > *one_hour* **then**
4: Assign new Session-ID for that log entry
5: **else**
6: Assign the old Session-ID.
7: **end if**
8: **end for**

Chapter 6

Exploratory Data Analysis

After having better understanding of the dataset in section 5.1 we can explore the data to understand the behavior of the dataset and its variables. As the Hotel Y has online visitors from all over the glob, it's interesting to analyze the data based on locations. In this section 6.1 we will analyze our data based on geographic locations.

6.1 Geographical Analysis

Geographical analysis is used when any direct or indirect reference to the geographic location is available in the data. Location here can be different countries, states or cities. Direct reference means the data contains name of the location or it may contains longitude and latitude of a particular location. However, indirect references can be *ip-address* which can reference location of online users.

6.1.1 Most Visitors By Country

As such 57% which is more than half of the total online visitors to the site are from the Hong Kong. The next three countries are America, Australia, and England. Hotel Y has got 2% visitors from all three countries named Singapore, Japan and Taiwan. As chart in the Fig. 6.1 mentions that, there are 16% visitors are from other countries around the world. This results can assist in marketing, promotions and can also provide insight into possible locations for content distribution network(CDN) expansion or Server placement to better serve visitors physically distant from the HKG server.

6.1.2 Most visitors by state

Now, lets explore the state of a particular country and check out the proportion of online visitors from particular state. As it is noted from section 6.1.1 that Hotel Y has highest visitors from Hong Kong and we have got similar results where highest users are from

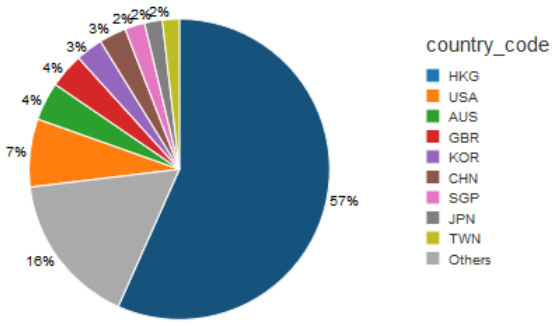

Figure 6.1: Most Visitors By Country

Kowloon City, Hong Kong. Proportion of other major states are shown in Fig. 6.2. Hence, these results can be used in decision making for marketing and promotions in a particular state.

6.1.3 Cities making most requests in top four countries

We analyzed from these top four countries where the majority of requests were coming from within each city Fig. 6.3 represents that, The majority(77%) of Hong Kong visitors are from within the Central District. 28% (880k), 8% (240k), 5% (162k) of American visitors are from Mountain View, New York and Redmond respectively. The majority of U.K. visitors are from London, at 40% (346k). Moreover, 23% (240k), 15% (158k), and 12% (129k) of Australian visitors are from Sydney, Melbourne and Perth respectively.

6.1.4 Response Time By Country

Response times directly correlate to user engagement, and in fact user retention. Lowering load times in cities or countries could possibly lead to more user bookings compared to competitors. Extra analysis was performed for Response Times By Country. For example, the Vatican visited the website one 1 occasion, therefore skewing the results of the average times. Countries with very low hits have been Filtered out from the results.

6.2 Server Load and Maintenance time

In this section, our major focus is on time when the server gets user request and the time to be taken by server to respond the request. Here, the term Server Load means the time when server gets busy and maintenance time leads to the suitable time when server maintenance work should be carried out.

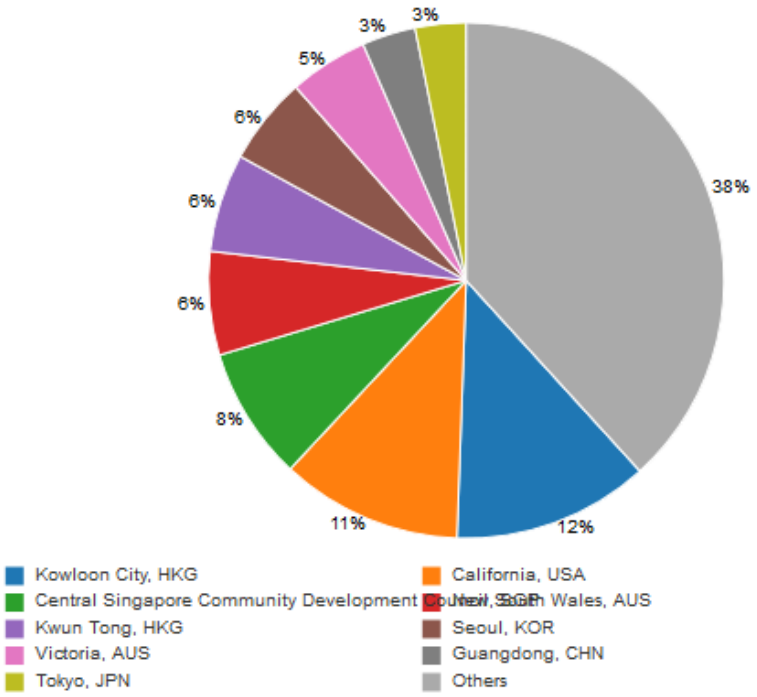

Figure 6.2: Most Visitors By States

6.2.1 Website Requests By Month

To measure the load of Hotel Y Website Server, user's web requests to the web-server can be calculated over days, weeks, months or years depends on the requirements. Here for our analysis, web requests are calculated over months. Fig. 6.5 highlights the line graph pointing to the number of hits in a particular month of the given duration. It is clearly seen that Hotel Y has got highest website visits in the month of March, 2015 while the least visits observed in month of August, 2014. However, the overall trend of online users is increased from August,2014 to August,2015 which is showing considerable future growth for Hotel Y.

As we have seen in the Fig. 6.5, month of March,2105 has the highest hits, we are more interested to see the results for this particular month to see hits from various countries in this month. Fig. 6.6 shows percentages of total hits by country.

6.2.2 Website Requests By Hours

Similar to the section 6.2.1 results, average web requests can be calculated on hourly bases. The Fig. 6.7 depicts the line graph pointing out the average number of requests during 24 hours of the day.

It can observed from the Fig. 6.7 that, during 3 : 00AM to 10 : 00AM, user's are

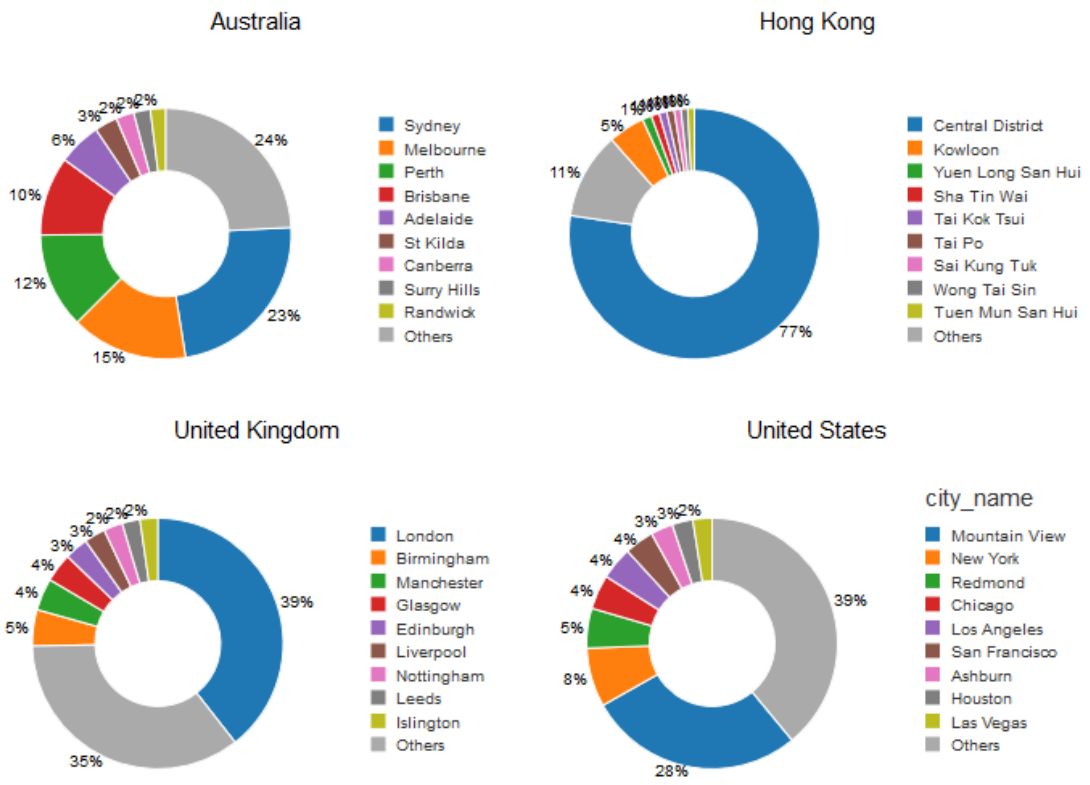

Figure 6.3: Most Visitors By States

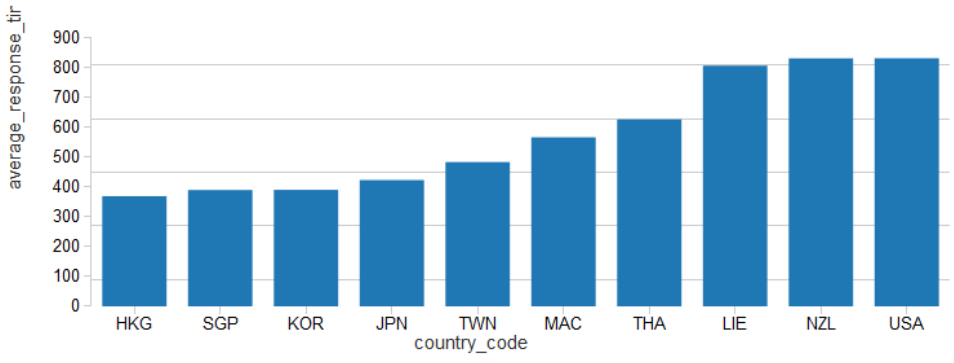

Figure 6.4: Fastest 10 countries with significant traffic

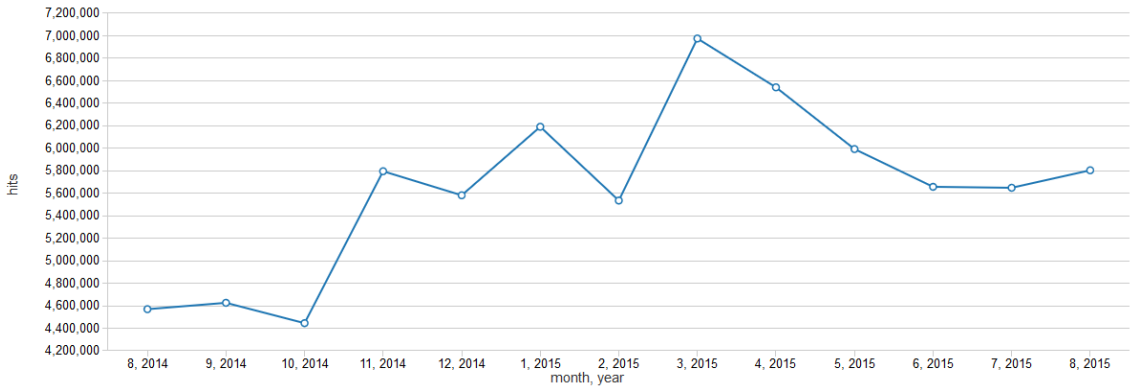

Figure 6.5: Website Requests By Month

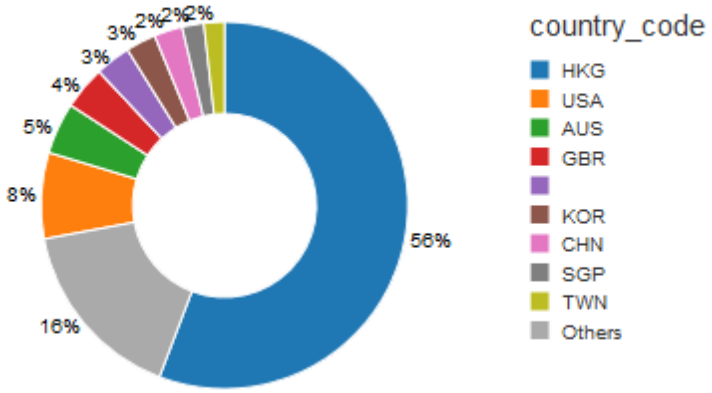

Figure 6.6: Website Requests in the month of March,2015 by top countries

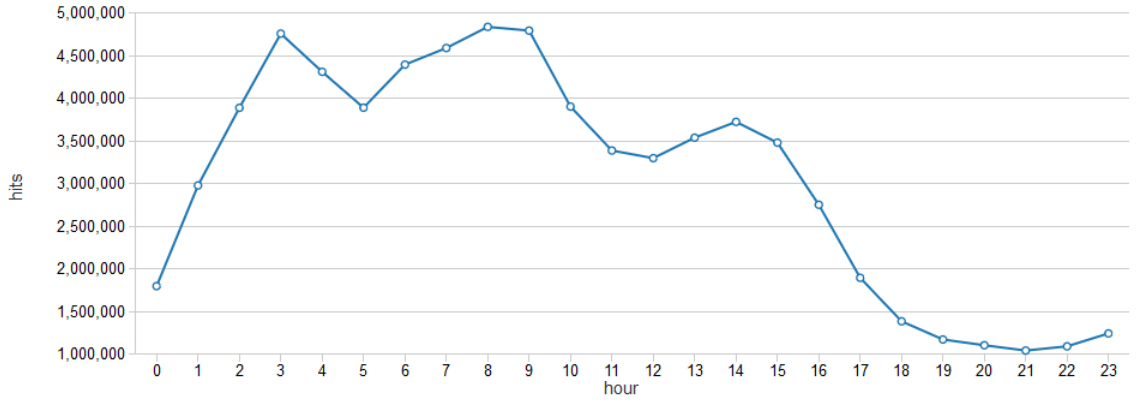

Figure 6.7: Website Requests By Hours

accessing Hotel Y website highest throughout the day. So we are more interested in this particular duration of the day. Fig. 6.8 shows hits by countries during 3 : 00AM to 10 : 00AM where Hong Kong, USA and Australia are top three countries.

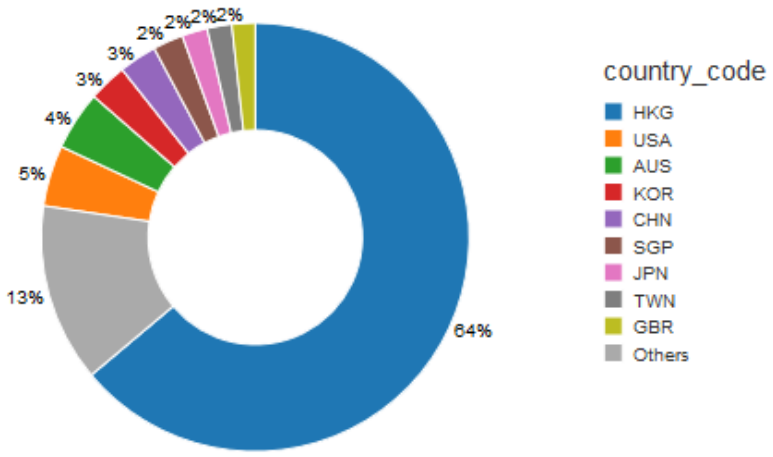

Figure 6.8: Website Requests between 3:00-10:00am by countries

6.2.3 Server Errors

The number of errors being made are increasing with a steady incline, from 500 in August 2014 up to 4,750 a year later. This is plotted in Fig. 6.10. This 89% increase in errors may also be due to an increase in requests (Fig. 6.5). It is notable that a majority of the requests 15k errors are caused by humans in Hong Kong, whereas 7k are by bots from the United States. While Fig. 6.9 shows internal server errors occurred during hours of the day. It is observed that during 7am to 2pm, server has got highest internal errors as from the Fig. 6.7, it is evident that server has got highest requests during this time duration.

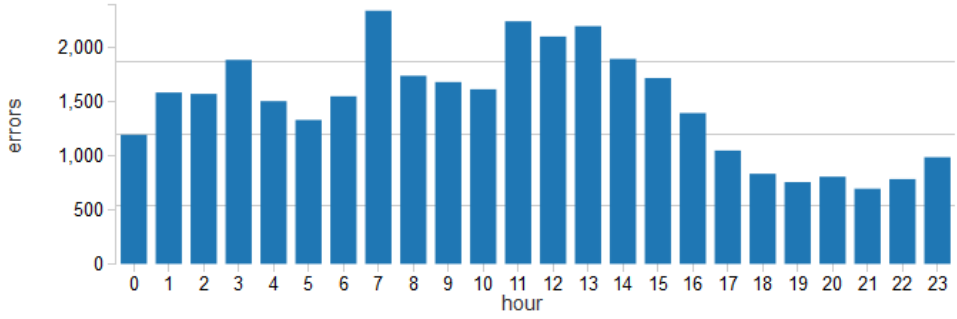

Figure 6.9: Internal Server errors by hours

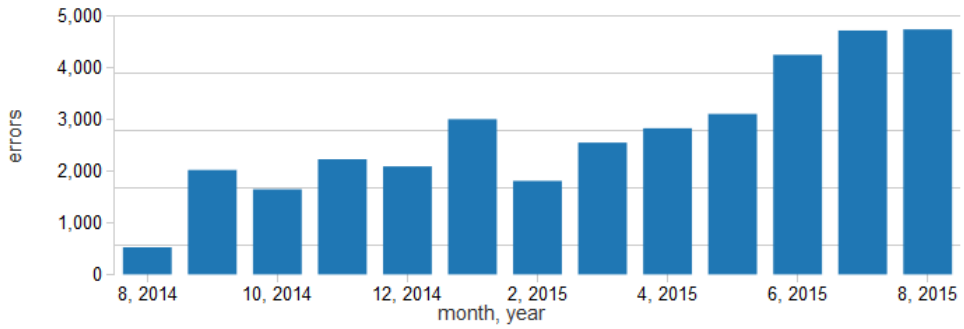

Figure 6.10: Internal Server errors by months

Chapter 7

Frequent Pattern Mining

Frequent pattern Mining has always considered as one of the significant data analysis task in the field of data analysis. In our scenario, It has much more significance compare to other supervised learning methods. Web visit of a particular web page of Hotel Website is considered as one visit and then frequency of that page-visit can be counted and taken into consideration. Similarly, we can consider visit of web-page B after web-page A and how frequently that happened. So our aim of using this algorithm here is to identify frequent web-page visits of users.

7.1 Data Preprocessing

7.1.1 Feature Selection

As per the data dictionary, we have total 14 features in the log files. There are some of the features that we need to ignore and select all of those features which are needed to satisfy our analysis objectives. Here, below five features must be selected for our web page pattern analysis.

1. date
2. time
3. cs_ip
4. cs_uri_stem
5. user_agent

7.1.2 Data Reduction: Sessionization

A web session is defined as a single session of a particular user at a given moment of time. It consists of what pages were visited within that period of time, as determined by the requests made to the server in this period.

To extrapolate meaningful data, the definition of what a user session was needed to be determined. This is because we need to find patterns made in one particular user session; when a user visits the website in one sitting, what pages do they visit and in what order?

To do this, we assume that a session is made up of the following factors within a series of web requests:

1. The clients IP address must be the same,
2. The clients user agent string must be the same,
3. The timestamp of the session is grouped in the same day, and
4. The timestamp of the session is within the same hour.

We group every request using these four key factors, using a hash as a delimiter, into the field session identifier. Using this identifier, we can group up a series of requests into one particular session. For example, a sample request is made to the server:

- The c_ip field is 10.20.30.40,
- The user agent field is Safari,
- The timestamp field is Wed Jan 0104 : 20 : 00 GMT 2015,

Therefore the session identifier would be: 10.20.30.40#Safari#01012015#04

7.2 Contrast Analysis

Contrast analysis has been accomplished using frequent pattern mining for web page access patterns. The term Contrast analysis means analyzing the difference in the behavior between two sets of users. In this section, contrast between Internal and External visitors in Sec. 7.2.1 7.2.2, Hong Kong and International visitors in Sec. 7.2.3 7.2.4, Visitors from top three countries in Sec. 7.2.3 7.2.5 7.2.6 have been briefly explained using graphs.

7.2.1 Internal Requests

Fig. 7.1 visualizes the frequency patterns found from internal visitors from the website. Table 7.1 further extrapolates this data and identifies the frequency pattern values. We have identified the most common frequencies of user sessions within the hotels network, ordered from most to least common. The most common internal visitor pattern identified indicates that current visitors are looking to extend their stay or visit the hotel again at a later date. It may also be possible that these are not guestsbthey may be potential guests who are not checked in, have walked into the hotel, connected to the lobbys guest network and are interested in staying at the hotel, thereby using the internal 10 network to find out more about the rooms, dining experiences, offers available and about the hotels facilities.

Table 7.1: Internal Web Requests

Sequence	**From**	→	**To**	**Frequency**
0	about the hotel	→	rooms	123
0	rooms	→	offers	123
1	about the hotel	→	dining	122
1	dining	→	offers	122
2	facilities	→	about the hotel	103
2	about the hotel	→	offers	103
3	rooms	→	dining	102
3	dining	→	offers	102
4	facilities	→	about the hotel	101
4	about the hotel	→	rooms	101
5	facilities	→	rooms	99
5	rooms	→	offers	99
6	facilities	→	dining	96
6	dining	→	offers	96
7	about the hotel	→	rooms	96
7	rooms	→	dining	96
8	above and beyond	→	dining	94
8	dining	→	offers	94

Visitors connected to the internal network typically end their session on the Offers pagethis information is useful and we suggest that the hotel priorities the display of offers on its website more predominantly. This would attract more customers by presenting more offers when they are connected to the hotels network.

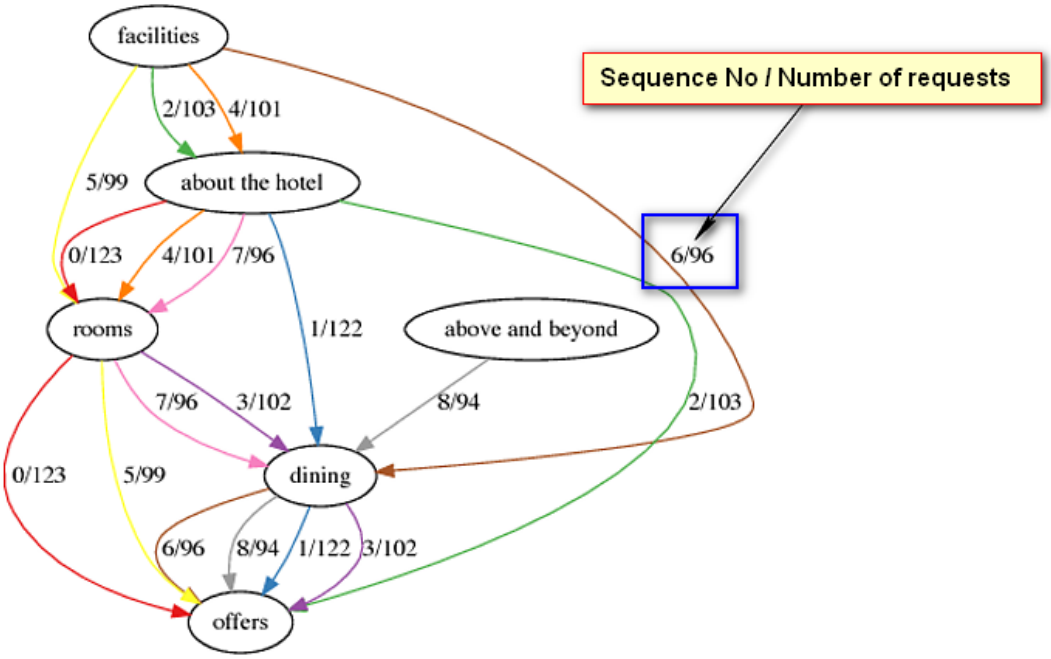

Figure 7.1: Navigation Patterns: Internal requests

7.2.2 External Requests

Fig. 7.2 visualizes the frequency patterns found from external visitors. The external IP request sources are visitors to the TULIP hotel who are accessing the site outside of the TULIP Hotel network. Most visitors will fall into this category. We have identified the following patterns in external visitors user sessions:

1. 'Above and Beyond' → 'Facilities' → 'Rooms',
2. 'Above and Beyond' → 'Offers' → 'Dining',
3. 'Facilities' → 'Rooms' → 'Offers',
4. 'Facilities' → 'Offers' → 'Dining',
5. 'Facilities' → 'Rooms' → 'Dining',
6. 'Above and Beyond' → 'Facilities' → 'Dining',

We have identified that most common external user sessions navigate between a mix of 'Above and Beyond, 'Facilities, 'Rooms and 'Dining; this highlights a common interest of these users. Patterns also highly suggest that many users navigate to the 'Offers page, most commonly from 'Above and Beyond followed by 'Rooms and finally 'Facilities.

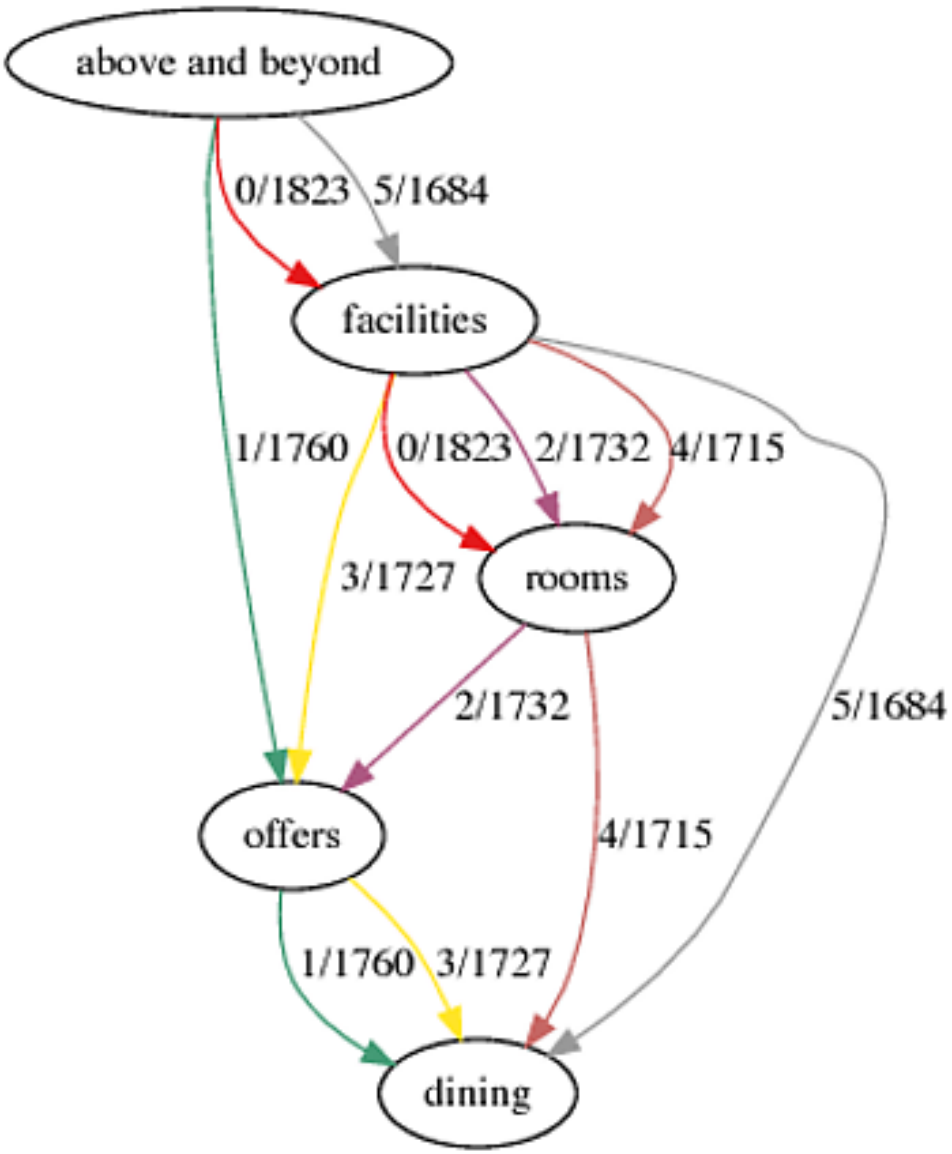

Figure 7.2: Navigation Patterns: External requests

We propose that these patterns could be implemented into the UI of the website to highlight how room information is highly sought afterthis could be made more prominent on the website, most particularly under the Facilities page, which will help users find the information they typically look for after reviewing what facilities the hotel has to offer. User sessions for this visitor category typically end at the 'Offers, 'Rooms and 'Dining; we suggest these pages should be highlighted as closely as possible in the homepage as it shows what pages and information is most sought after by users once they have visited the website.

7.2.3 Requests from users in Hong Kong

Fig. 7.3 visualizes the user sessions of Hong Kong visitors. The top three patterns identified by Hong Kong visitors are:

1. 'Above and Beyond' → 'Offers' → 'Dinning',
2. 'Facilities' → 'Offers' → 'Dining',
3. 'About the Hotel' → 'Offers' → 'Dining',

7.2.4 Requests from users outside Hong Kong

Fig. 7.4 visualizes the user sessions for visitors outside the Hong Kong . The top five patterns identified by international visitors are:

1. 'Above and Beyond' → 'Facilities' → 'Rooms',
2. 'Facilities' → 'Rooms' → 'Offers',
3. 'Above and Beyond' → 'Rooms' → 'Offers',
4. 'Facilities' → 'Dinning' → 'Rooms',
5. 'About the Hotel' → 'Rooms' → 'Offers',

7.2.5 Requests from USA users

Fig. 7.5 visualizes the user sessions for visitors from USA . Table 7.2 further explains this data and identifies the frequency pattern values. The top five patterns identified by USA visitors are:

1. 'Rooms' → 'Offers' → 'About the Hotel',

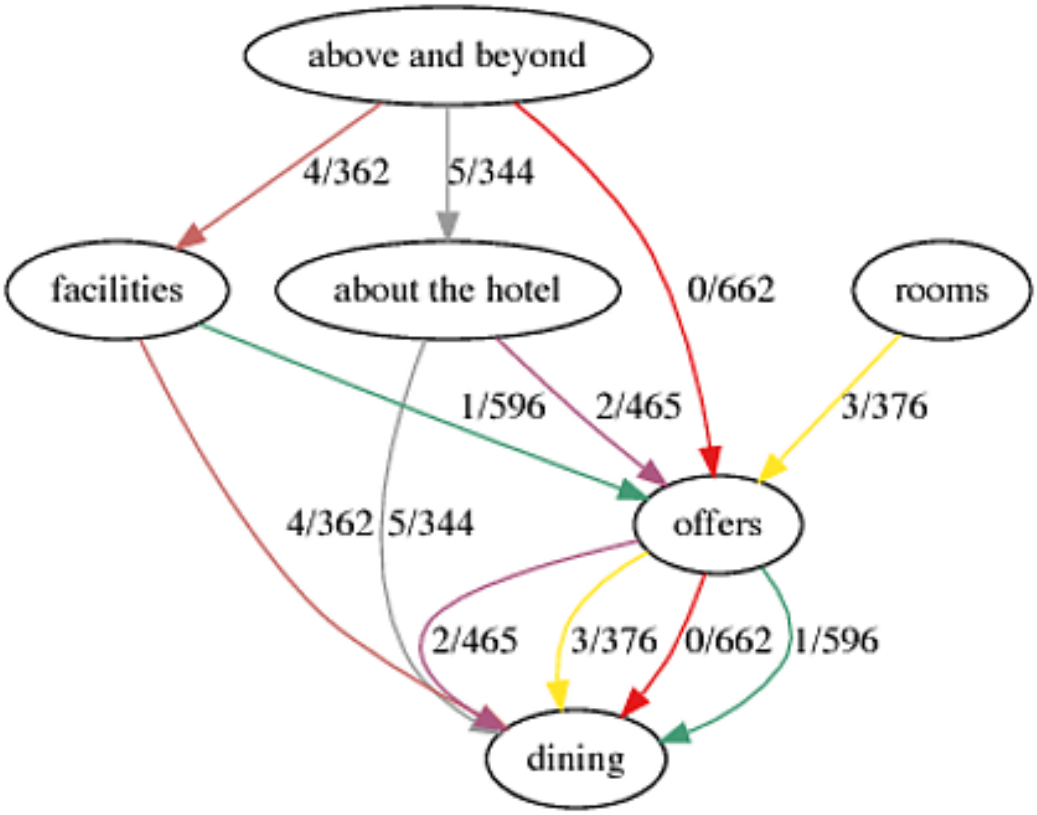

Figure 7.3: Navigation Patterns: Hong Kong requests

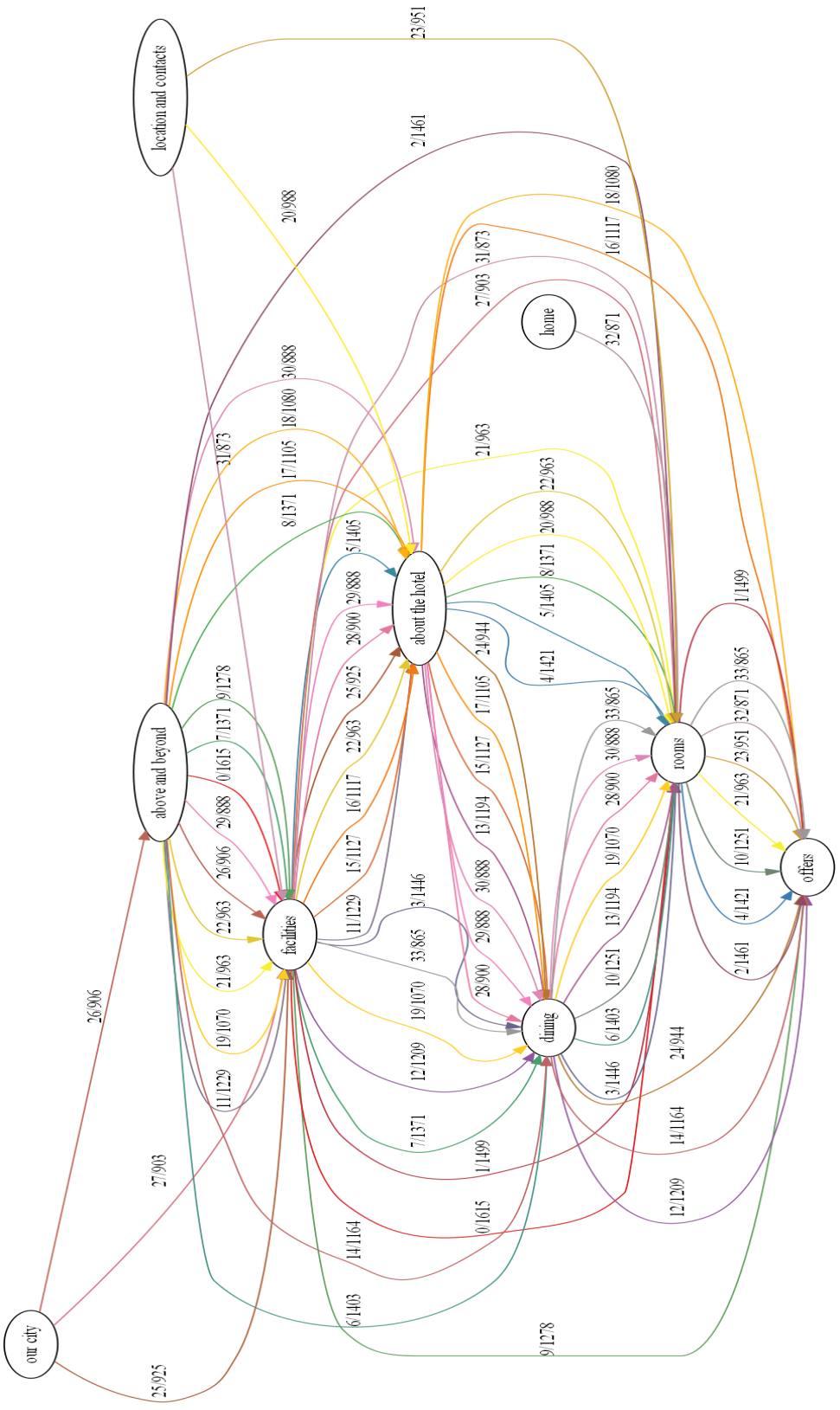

Figure 7.4: Navigation Patterns: Outside of Hong Kong requests

2. 'Above and Beyond' → 'Rooms' → 'About the Hotel',
3. 'Above and Beyond' → 'Rooms' → 'Offers',
4. 'Facilities' → 'Rooms' → 'About the Hotel',
5. 'Above and Beyond' → 'Offers' → 'Rooms',

Table 7.2: USA Web Requests

Sequence	**From**	→	**To**	**Frequency**
0	rooms	→	offers	330
0	offers	→	about the hotel	330
1	above and beyond	→	rooms	329
1	rooms	→	about the hotel	329
2	above and beyond	→	rooms	299
2	rooms	→	offers	299
3	facilities	→	rooms	281
3	rooms	→	about the hotel	281
4	above and beyond	→	offers	273
4	offers	→	about the hotel	273
5	above and beyond	→	facilities	272
5	facilities	→	rooms	272
6	rooms	→	dining	268
6	dining	→	offers	268
7	our city	→	facilities	260
7	facilities	→	about the hotel	260

7.2.6 Requests from Australian users

Fig. 7.6 visualizes the user sessions for visitors from USA . Table 7.3 further explains this data and identifies the frequency pattern values. The top five patterns identified by USA visitors are:

1. 'Rooms' → 'Offers' → 'About the Hotel',
2. 'Above and Beyond' → 'Rooms' → 'About the Hotel',
3. 'Above and Beyond' → 'Rooms' → 'Offers',
4. 'Facilities' → 'Rooms' → 'About the Hotel',
5. 'Above and Beyond' → 'Offers' → 'Rooms',

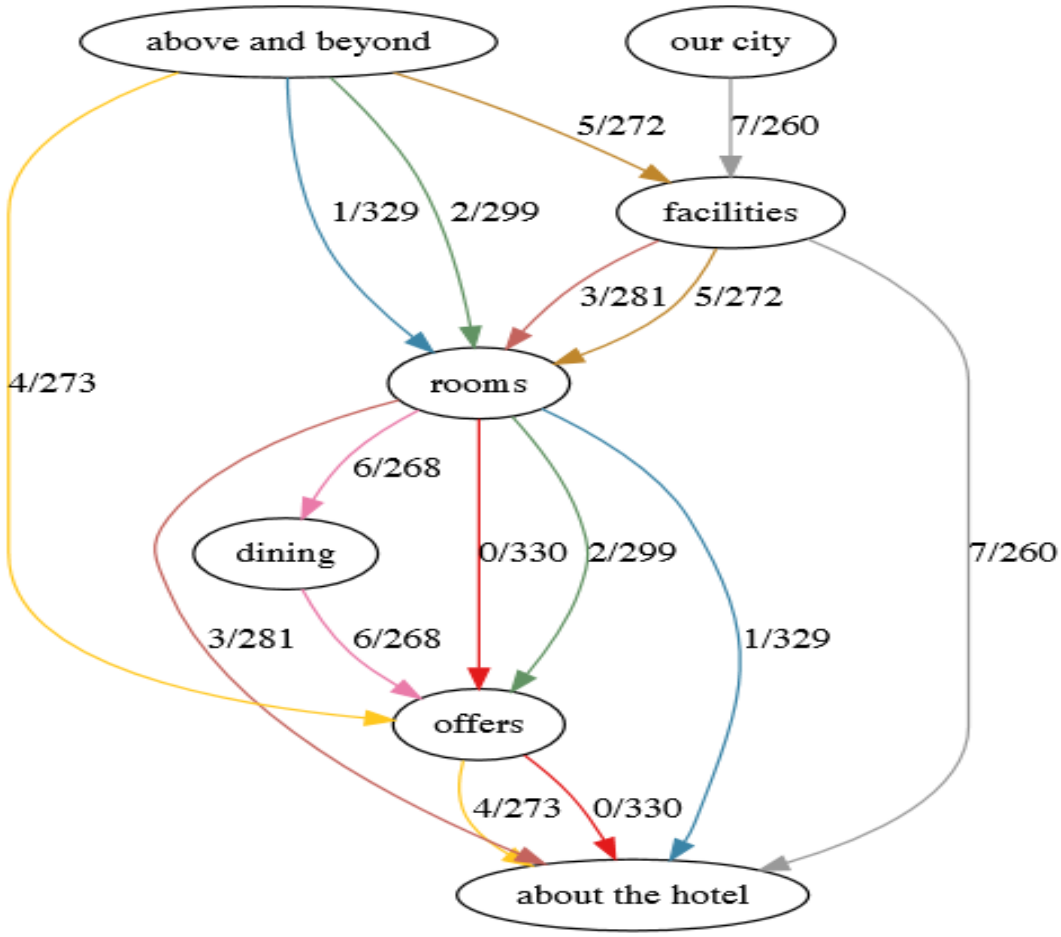

Figure 7.5: Navigation Patterns: USA requests

Table 7.3: Australia Web Requests

Sequence	**From**	→	**To**	**Frequency**
0	above and beyond	→	facilities	152
0	facilities	→	rooms	152
1	above and beyond	→	rooms	151
1	rooms	→	offers	151
2	facilities	→	rooms	141
2	rooms	→	offers	141
3	about the hotel	→	rooms	122
3	about the hotel	→	offers	122
4	about the hotel	→	facilities	120
4	facilities	→	rooms	120
5	above and beyond	→	dining	120
5	dining	→	rooms	120
6	above and beyond	→	facilities	115
6	facilities	→	offers	115
7	facilities	→	dining	107
7	dining	→	rooms	107
8	above and beyond	→	facilities	105
8	facilities	→	dining	105
9	dining	→	rooms	103
9	rooms	→	offers	103
10	about the hotel	→	above and beyond	100
10	above and beyond	→	facilities	100
11	above and beyond	→	dining	100
11	dining	→	offers	100
12	about the hotel	→	above and beyond	97
12	above and beyond	→	rooms	97

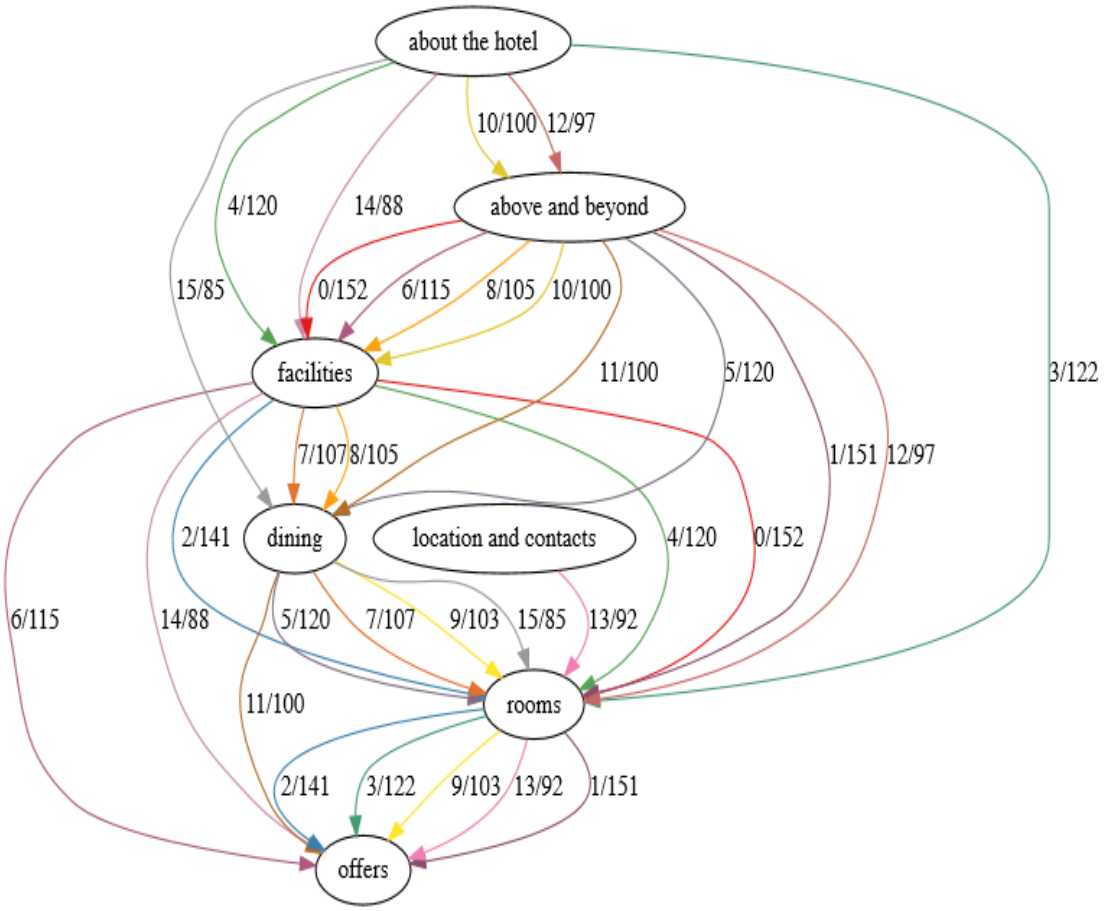

Figure 7.6: Navigation Patterns: AU(Australian) Web Requests

Chapter 8

Conclusion

This Thesis is the brief study of the problem of identifying customer behavior by discovering hidden patterns in the large volume of log data collected by the web server. Here the analysis of FP-Growth algorithm justified that it is more effective than Apriori algorithm. Also it takes less time and better in performance [2]. All in all to conclude this report, we can focus more on applications of this web log analysis. As the main objective we will achieve here is to give strong suggestions to website designers and administrators to ensure adequate capacity of web server and the outputs of frequent pattern mining will help them to identify web pages to be redesigned and improve its loading time to attract more visitors and hence in this way to increase the overall user experience as well as customer satisfaction.

8.1 Proposed Suggestions

After analyzing all the results, this thesis can propose some useful suggestions to the Hotel Y which Hotel Y can consider while decision making such as advertising and marketing, improving their online user's experience etc. As this analysis reveled customer's visiting behavior for various sets of customers, proposed suggestions are usefulness of those insights and includes followings.

- From Sec. 7.2.1, Visitors connected to the internal network of Hotel Y, mostly end their session on the 'Offers' page which is useful for suggesting that the hotel should prioritizes the display of offers page on the website more predominantly. This would help in attracting more customers by providing more offers while being connected with the hotels network.
- From Se. 7.2.2, external visitors from all over the world typically ends their sessions at the 'Offers', 'Rooms' or 'Dining' pages. So it is strongly suggested that these pages should me linked and placed as closely as possible in the homepage of the Hotel website.

- From Sec. 7.2.3, we can suggest that hotel Y should do more promotions and offers on 'Offers' web-page as this promotional campaign can target more Hong Kong visitors.
- From Sec. 7.2.5, it is identified that most of the American users tend to end their web session by accessing the 'About the Hotel' page. So the suggestion would be to make sure that 'About the Hotel' page is capable of providing sufficient information that the visitors are looking for.
- From Sec. 7.2.6, It is observed that Australian visitors tend to review the 'Facilities' page before visiting the Hotels 'Rooms' and 'Offers' page. So it can be suggested that the order of displaying these pages effectively might be a good option to explore for Australian visitors.

8.2 Future Work

As mentioned in the section 4.1, web log files are an extensive source of stream data and hence Use of these log files is significant and can not be limited to identifying frequent patterns. The extensions of this sequential frequent pattern mining can be obtained by combing with server error detection which means that what particular access pattern ended with an error. Let's say access patten like, $Home \rightarrow Offers \rightarrow Rooms \rightarrow Facilities \rightarrow 404$. In this case, Facilities page leads to server error, which means that facilities page has some technical issue and needs to be solved. Similarly, it can be combined with payment gateway identification which will tell us what specific access path takes the user to the payment page. For example access path is, $Home \rightarrow Offers \rightarrow Rooms \rightarrow Booking \rightarrow Payment$. In this case, user has started from home page and ended by payment for room booking in the hotel. Such patterns are so helpful in increasing potential customers.

Bibliography

[1] Soumen Chakrabarti. Data mining for hypertext: A tutorial survey. *ACM SIGKDD Explorations Newsletter*, 1(2):1–11, 2000.

[2] K Dharmaraajan and MA Dorairangaswamy. Analysis of fp-growth and apriori algorithms on pattern discovery from weblog data. In *Advances in Computer Applications (ICACA), IEEE International Conference on*, pages 170–174. IEEE, 2016.

[3] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. Hypertext transfer protocol–http/1.1. Technical report, 1999.

[4] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM sigmod record*, volume 29, pages 1–12. ACM, 2000.

[5] Renáta Iváncsy and István Vajk. Frequent pattern mining in web log data. *Acta Polytechnica Hungarica*, 3(1):77–90, 2006.

[6] Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The web as a graph: measurements, models, and methods. *Computing and combinatorics*, pages 1–17, 1999.

[7] Naga Lakshmi, Raja Sekhara Rao, and Sai Satyanarayana Reddy. An overview of preprocessing on web log data for web usage analysis. *IJITEE, ISSN*, pages 2278–3075, 2013.

[8] Rob Law and Billy Bai. *Website Development and Evaluations in Tourism: A Retrospective Analysis*, pages 1–12. Springer Vienna, Vienna, 2006.

[9] Rob Law, Shanshan Qi, and Dimitrios Buhalis. Progress in tourism management: A review of website evaluation in tourism research. *Tourism Management*, 31(3):297 – 313, 2010.

[10] Rosanna Leung and Rob Law. *Analyzing a Hotel Website's Access Paths*, pages 255–266. Springer Vienna, Vienna, 2008.

[11] James N.K. Liu and Elaine Yulan Zhang. An investigation of factors affecting customer selection of online hotel booking channels. *International Journal of Hospitality Management*, 39(Supplement C):71 – 83, 2014.

[12] Mingte Lu and Winglok Yeung. A framework for effective commercial web application development. *Internet Research*, 8(2):166–173, 1998.

[13] Microsoft. W3c extended log file format, 2017.

[14] G Neelima and Sireesha Rodda. An overview on web usage mining. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*, pages 647–655. Springer, 2015.

[15] G Neelima and Sireesha Rodda. Predicting user behavior through sessions using the web log mining. In *Advances in Human Machine Interaction (HMI), 2016 International Conference on*, pages 1–5. IEEE, 2016.

[16] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu. Web-page recommendation based on web usage and domain knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2574–2587, 2014.

[17] Charles Ruel Novabos, Aura Matias, and Miguela Mena. How good is this destination website: A user-centered evaluation of provincial tourism websites. *Procedia Manufacturing*, 3(Supplement C):3478 – 3485, 2015. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

[18] Rajinder Singh Rao and Jyoti Arora. A survey on methods used in web usage mining. 2017.

[19] R Shanthi and Dr SP Rajagopalan. An efficient web mining algorithm to mine web log information. *IJIRCCE*, 1(7), 2013.

[20] Liang Wang, Rob Law, Basak Denizci Guillet, Kam Hung, and Davis Ka Chio Fong. Impact of hotel website quality on online booking intentions: etrust as a mediator. *International Journal of Hospitality Management*, 47(Supplement C):108 – 115, 2015.

[21] Qiang Yang, Charles X. Ling, and Jianfeng Gao. *Mining Web Logs for Actionable Knowledge*, pages 169–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[22] Tom Au Yeung and Rob Law. Extending the modified heuristic usability evaluation technique to chain and independent hotel websites. *International Journal of Hospitality Management*, 23(3):307 – 313, 2004.