Analysis of Students' Web Browsing Behaviours Using Data Mining at a Campus Network

Murizah Kassim, Muhammad Ahlami Ashraf Roslan
School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 UiTM Shah Alam, Selangor, MALAYSIA
murizah@uitm.edu.my

Abstract: Analytics provides insight to people based on the analytics of past usage by using techniques such as statistics, data mining, machine learning and artificial intelligence. Lack of monitoring system of browsing causes low engagements that reduce the growth of certain businesses caused by unnecessary browsing for students learning time. This paper presents an analysis on browsing behavior that classifies browsed words followed their ethical word-groups browsing. An Analytic platform is created as a monitoring system of browsing behavior. Data mining, indexing and classification method are used in this research as data is the essential key of creating a predictive model and four types of ethical groups have been filtered based on the browsing behaviors. The browsed words are categorized into four types of browsing called queries, applications, social media, Campus-related sites. The research method uses software tools and data mining process on the browsing data and analytics is presented on the development of the dashboard mainly using the R programming language. Few unethical words using the indexing method are generated in analytic graphs based on the type of browsing versus time. Data collected from the browsing behaviors of students'analysis taken from browsing database of personal computer and laboratory computer in the campus network. The result shows that othercategories are the highest categories which reached79.6% for personals' computer browsing compared to72.4% browsing at the laboratory computers. It is identified that about 21% of the browsing behavior was filtered during the data mined processed. The other category is still on the research portfolio where these libraries must be filtered in detail to identify whether they are learning or non-learning activities. This research is significant in that helps to increase the effectiveness of suggestions applications, optimize the internet usage by blocking unnecessary words or webpages, and even campus guide systems by monitoring the surrounding browsing behavior of the students' usages of the campus network computer labs.

Keywords: Browsing behaviors, Data Mining, Campus Network, R programming, Data Cleaning, Classification.

1. Introduction

The web has changed dramatically over the last two decades with online information was published scholarly and freely. Technology has changed where the internet is increasingly used to gain knowledge and understanding of a topic. This knowledge is often acquired by browsing on the web and some factors that have influenced critical internet use have become social. The increased usage of the internet has shown the increasing numbers of words used for browsing purposes(Kaushik & Jones, 2018). Web browsing behaviour is studied has been done using web usage mining with its applications in adaptive and personalized systems. Current web browsers allow opening multiple web pages at once and switching between them called parallel browsing. Client-side observations are performed which poses a challenge in attracting enough participants to capture browsing behaviour. A study described an experiment on logging the parallel browsing behaviour such as in an adaptive web-based educational system and on the open Web. Some are using the educational system as a tool for recruiting and motivating the participants(Labaj & Bieliková, 2015). A study of consumer browsing behaviours has been utilized to perform the tagging task automatically. A proposed novel graphical model and the development of a new algorithm have been done to properly integrate both browsing behaviours and content information for automatic tagging. The research studied has contributed to machine learning techniques (Liu & Liu, 2021). Subjective reports indicated that some online users were becoming addicted to the Internet in much the same way that others became addicted to drugs or alcohol, which resulted in academic, social, and occupational impairment. However, research among sociologists, psychologists, or psychiatrists has not formally identified the addictive use of the Internet as a problematic behaviour. Every activity that we do on the internet will leave a print or in a computer's term called history. This database of history can be used for a better purpose which one of them is using for predictive analytics. A lot of data is needed to create a solid projection of analytic. The main problem for this research was to have clean data that is not disturbed or cleaned by certain features such as incognito mode. It is a fact that the user will use incognito mode to browse unnecessary and unethical things to hide from leaving a cyber footprint. This private browsing was made for different goals but is used for a different purpose by the user as mentions in a study(Bursztein et al., 2010).

Web browsing involved prescriptive perspectives on the correct usage of words like unique, geographical information, labelling, and many more. This research focuses on providing analytics of ethical browsing of total browsing and personalize browsing and projecting a prediction for the user based on word categories such as queries, social media, applications, and unethical words. To complete this research, there are several bits of knowledge that need to be gained such as the knowledge of data science, programming language, analytics, and the most important part is valid data. Data science is about the way of blending a multidisciplinary of data inference, algorithm development, and technology to solve analytically complex problems. Possibly the most closely related concept to data science is data mining, the actual extraction of knowledge from data via technologies that incorporate these principles(Provost & Fawcett, 2013). Mastering the programming language of Python and R programming are also compulsory to complete this research. For data analysis and interactive, exploratory computing and data visualization, Python will inevitably draw comparisons with other domain-specific open source and commercial programming languages and tools in wide use such as R, MATLAB, SAS, Stata, and others(Surya Gunawan et al., 2020). A research has used R programming as its platform because of its advantage in data cleaning and provides lots of packages for data analysis and graphics to plot the results. Software tools named R packages for scripting are used and MYSQL as a database is performed to store and process data using the cloud services. The data on cloud services can be always accessible from any machine and. Data are reliably backed up and data retrieval times are comparable to a public cloud storage service(Sharma et al., 2020). The data is divided into three groups which are personal computers and laboratory computers to show the difference between browsing behaviour in different situations. Although aggregating user-generated data is useful, finding patterns in it can help answer questions about the way the world works(Gao et al., 2021). This research filters the words fetched from the words that are browsed by people from different situations.

2. Literature Review

Analytics is a powerful tool that has been used for ages by categorizing the variables in different groups.Most browsers record all history of visited web pages for future revisitation. One of the reasons is that the records are displayed at once as a single list, which may devastate the users. A research has proposed a predictive model to decide whether a web page will be revisited in the future based on a particular visit. The model can be used to filter web records so that only web pages that may be re-visited are presented. According to our evaluation, the model is considerably effective(Toba et al., 2020). Predictive analytics of ethical words for web browsing is about categorizing words that were browsed by the user in different browsing situations such as personal computers and laboratory in groups such as social media, queries, applications, and Campus related sites. A study on self-similarity Hurst parameter estimation with rescaled range method on IP-based campus internet traffic has been done to analyse the repeated data of internet data (Kassim et al., 2017). This data will generate a model that can store user's rates of groups that they will browse in the future. There are lots of processes involved in creating a solid model such as data mining, data cleaning, classification, and generating an equation. Using an R programming language, all these processes can be done all in one platform. However, a high amount of data and authentic data is essential to create a very solid and stable model. Data mining is the first important step to create a solid model. A sample is simply a subset of the population, preferably a representative subset(Astika Saputra et al., 2020). Data mining is a process of extracting data and information related to the research. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information of knowledge from the rapidly growing volumes of digital data(Abramson & Aha, 2013). This process turns raw data into useful data using statistical ways. This method is the way everything evolves into something better by learning from the past. There are lots of types of data mining such as relational databases, data warehouses, abject-oriented and more, but for this research, text mining and web mining technique are used to gain data.

Machine learning can be defined as a set of methods that can automatically detect patterns in data and using it to predict future data or to perform other kinds of decision-making under certainty. Machine learning is usually divided into two main types which are supervised and unsupervised. This research is using the supervised learning approach which needs a mapping of input to create output which is the predictions. Supervised learning which also called inductive machine learning is a process that learns from a set of rules that creates a classifier that can be used to predict new data(Kotsiantis, 2007). The classification method is used for this research as it is more suitable to label the words based on the words library database. Why is the classification method being more suitable for this research? Based on a past study, classification is the prediction of labels while regression is predictions of quantity. The browsed words are grouped by the algorithm and put inside the word indexing table showed the frequent use type of ethical words by a person. The accuracy of a classification predictive model is based on the percentage of correctly classified examples out of all predictions made. Supervised machine learning was used in most practical machine learning. It is the construction of algorithms that can produce hypotheses and general patterns by using externally supplied instances to predict the fate of future instances(Das et al., 2015). The classification method is used to classify the type of ethical words that are being browsed. The learning process in machine learning begins with observations or data. For example, observations or data from direct experience, or instruction to look for patterns in data and make better decisions based on inputs or examples provided. Indexing is a method of classifying words into their categories. This method has been widely used especially for filtering and blocking purposes. Indexing is known as a data structure method that allows quick retrieval of records from a database file. It is like the search function in the computer where it filters all the string that is in the database and extracts it out. Indexing helps to create classifications of words that are then used to generate analytics as to the base of the model.

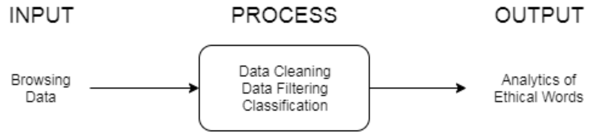
R is an open-source programming language that is widely used these days. It is mostly used by statisticians and data miners for developing statistical software and data analysis. It has a very friendly community that helps each other in developing their research. The advantage of R is that it has a network of File Transfer Protocols and web servers around the world that store identical, updated, versions of code and documentation for R. This helps to ease the user in terms of writing scripts. On top of that, it also provides graphics tools that can be run into the user interface via server using a package called Shiny.Shiny is one of the R packages that ease the process of building interactive web applications straight from R. Users can host standalone apps on a webpage or embed them in R. Shiny allows the user to interact, analyze and communicate with the data and analysis. This package helped increase the understanding of users towards the processed data and analytics. Database plays an important role in maintaining the model solid and better. MYSQL is open-source relational database management. It can either be stored inline via the cloud or via localhost. MySQL enables users to blend the best of both relational and NoSQL technologies into solutions that reduce cost, risk, and complexity including in-memory computing for real-time performance, integration with big data stores and frameworks, and online scaling and schema change. SQL can also be used as a tool for implementing the mining algorithm.Algorithms that allow computer programs to automatically improve through experience are considered machine learning. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation(Jalil et al., 2019; Jordan & Mitchell, 2015). An algorithm is a sequence or method of solving a problem or in other words, it is a series of instructions for carrying out an operation. It is mostly used in data processing, calculation, and other related computer and mathematical operations that suggests future opportunities to advance the research on data-efficiency in machine learning (Adadi, 2021). Four studies are related before this research with different methods and results. Table 1 shows the research gap in this research.

Reference Studies	**(Mathur et al., 2021)**	**(Zamfiroiu & Sbora, 2014)**	**(Lee et al., 2007)**	**(Diep et al., 2019)**
Type of Data	Words	Words	Words	Words
Method used	Classification	Statistical	The weightage (SMV)	Unsupervised
This study	Label words in two categories of positive or negative	Label words with the same meanings by their popularity of usage	Define bad words based on the number of bad words in a sentence.	Define the synonyms and antonyms of words

Table 1: Research gaps on data and method of indexing

3. Methodology

This research uses browsing data as the input and produces analytic as the output. The structure of this research is shown in Figure 3.This research only used software parts which consist of HTML, PHP, MySQL, and R scripting from the start until the end device. Each software has its scripting which has been written from scratch for this research. HTML and PHP scripting was written using Visual Studio Code while the database was set up using WampServer and the R script was written in RStudio. Every detail of this research has been shown and explained and Figure 4 shows the Software Flowchart for this research.



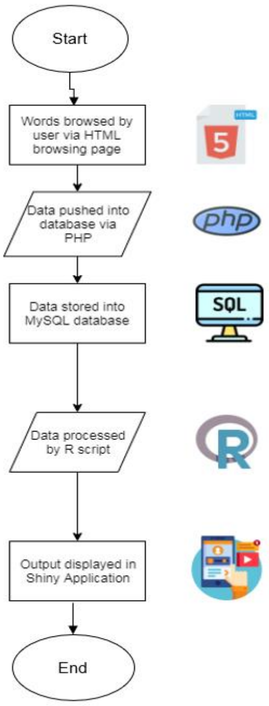

Start
Words browsed by user via HTML browsing page
Data pushed into database via PHP
Data stored into MySQL database
Data processed by R script
Output displayed in Shiny Application
End
php
SQL

Figure 1: Research Structure Diagram   Figure 2. System flowchart.

This research is using the classification and indexing method where the data are filtered based on words that are labelled in groups and are classified if it matches without uppercase or lowercase sensitive mode. Figure 3 shows the flowchart of this research from start to end.

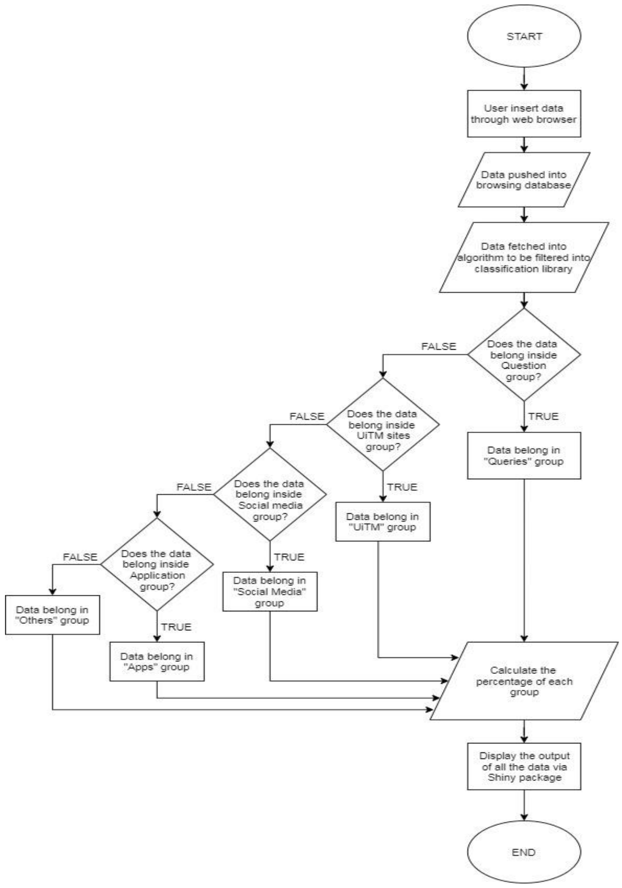
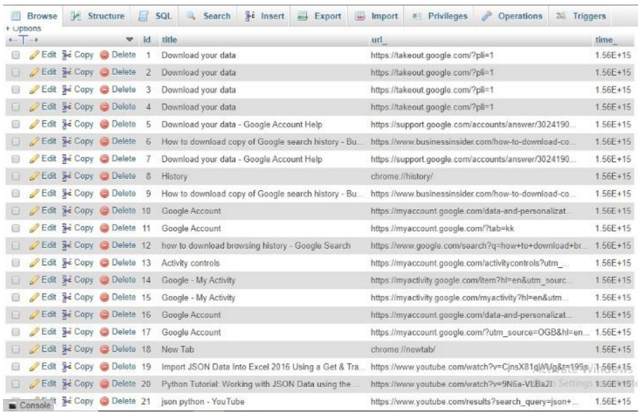

START
User insert data through web browser
Data pushed into browsing database
Data fetched into algorithm to be filtered into classification library
Does the data belong inside Question group?
FALSE
TRUE
Data belong in "Queries" group
Does the data belong inside UiTM sites group?
FALSE
TRUE
Data belong in "UiTM" group
Does the data belong inside Social media group?
FALSE
TRUE
Data belong in "Social Media" group
Does the data belong inside Application group?
FALSE
TRUE
Data belong in "Apps" group
Data belong in "Others" group
Calculate the percentage of each group
Display the output of all the data via Shiny package
END

Figure 2: System's Flowchart

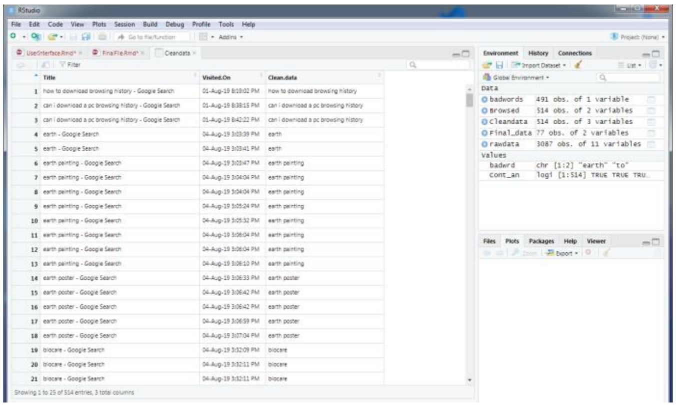
This research starts with the front page of the web browser. The first part of this research's code has all been written using the Visual Code Studio, a platform that supports lots of programming language and the markup language which is the Hypertext Markup Language, HTML, which is used in this research to make a sample of web browsing front webpage. The interface of the web browser will show the result using the Chromium browser. When the user enters the word for browsing, the words searched are automatically pushed into the MYSQL server and at the same time it shows the result of the browsed words. The words are pushed by the code that is attached to it written using Hypertext Pre-processor or PHP. The words browsed are recorded with the time and date details as shown in Figure 6.



	id	title	url_	time_
Edit Copy Delete	1	Download your data	https://takeout.google.com/?pli=1	1.56E+15
Edit Copy Delete	2	Download your data	https://takeout.google.com/?pli=1	1.56E+15
Edit Copy Delete	3	Download your data	https://takeout.google.com/?pli=1	1.56E+15
Edit Copy Delete	4	Download your data	https://takeout.google.com/?pli=1	1.56E+15
Edit Copy Delete	5	Download your data - Google Account Help	https://support.google.com/accounts/answer/3024190...	1.56E+15
Edit Copy Delete	6	How to download copy of Google search history - Bu...	https://www.businessinsider.com/how-to-download-co...	1.56E+15
Edit Copy Delete	7	Download your data - Google Account Help	https://support.google.com/accounts/answer/3024190...	1.56E+15
Edit Copy Delete	8	History	chrome://history/	1.56E+15
Edit Copy Delete	9	How to download copy of Google search history - Bu...	https://www.businessinsider.com/how-to-download-co...	1.56E+15
Edit Copy Delete	10	Google Account	https://myaccount.google.com/data-and-personalizat...	1.56E+15
Edit Copy Delete	11	Google Account	https://myaccount.google.com/?tab=kk	1.56E+15
Edit Copy Delete	12	how to download browsing history - Google Search	https://www.google.com/search?q=how+to+download+br...	1.56E+15
Edit Copy Delete	13	Activity controls	https://myaccount.google.com/activitycontrols?utm_...	1.56E+15
Edit Copy Delete	14	Google - My Activity	https://myactivity.google.com/item?hl=en&utm_sourc...	1.56E+15
Edit Copy Delete	15	Google - My Activity	https://myactivity.google.com/myactivity?hl=en&utm...	1.56E+15
Edit Copy Delete	16	Google Account	https://myaccount.google.com/data-and-personalizat...	1.56E+15
Edit Copy Delete	17	Google Account	https://myaccount.google.com/?utm_source=OGB&hl=en...	1.56E+15
Edit Copy Delete	18	New Tab	chrome://newtab/	1.56E+15
Edit Copy Delete	19	Import JSON Data Into Excel 2016 Using a Get & Tra...	https://www.youtube.com/watch?v=CjnsX81qWUg&t=199s	1.56E+15
Edit Copy Delete	20	Python Tutorial: Working with JSON Data using the ...	https://www.youtube.com/watch?v=9N6a-VLBa2I	1.56E+15
Copy Delete	21	json python - YouTube	https://www.youtube.com/results?search_query=json+...	1.56E+15

Figure 3: MySQL database research structure.

Moving on to the second part of the research, the data will then be fetched into another programming language that is called R. R programming is mainly used for analytics in this research. It will fetch the data from the MYSQL database and use the data to create an analysis that will show the browsing behaviour of the user. Figure 7 shows the data that has been fetched from the database and has gone through the cleaning process that is written in the R language.



	Title	Visited.On	Clean.data
1	how to download browsing history - Google Search	01-Aug-19 8:19:02 PM	how to download browsing history
2	can i download a pc browsing history - Google Search	01-Aug-19 8:38:15 PM	can i download a pc browsing history
3	can i download a pc browsing history - Google Search	01-Aug-19 8:42:22 PM	can i download a pc browsing history
4	earth - Google Search	04-Aug-19 3:03:39 PM	earth
5	earth - Google Search	04-Aug-19 3:03:41 PM	earth
6	earth painting - Google Search	04-Aug-19 3:03:47 PM	earth painting
7	earth painting - Google Search	04-Aug-19 3:04:04 PM	earth painting
8	earth painting - Google Search	04-Aug-19 3:04:04 PM	earth painting
9	earth painting - Google Search	04-Aug-19 3:05:24 PM	earth painting
10	earth painting - Google Search	04-Aug-19 3:05:32 PM	earth painting
11	earth painting - Google Search	04-Aug-19 3:06:04 PM	earth painting
12	earth painting - Google Search	04-Aug-19 3:06:04 PM	earth painting
13	earth painting - Google Search	04-Aug-19 3:06:10 PM	earth painting
14	earth poster - Google Search	04-Aug-19 3:06:33 PM	earth poster
15	earth poster - Google Search	04-Aug-19 3:06:42 PM	earth poster
16	earth poster - Google Search	04-Aug-19 3:06:42 PM	earth poster
17	earth poster - Google Search	04-Aug-19 3:06:59 PM	earth poster
18	earth poster - Google Search	04-Aug-19 3:07:04 PM	earth poster
19	biocare - Google Search	04-Aug-19 3:32:09 PM	biocare
20	biocare - Google Search	04-Aug-19 3:32:11 PM	biocare
21	biocare - Google Search	04-Aug-19 3:32:11 PM	biocare

Showing 1 to 25 of 514 entries, 3 total columns

```
Data
badwords    491 obs. of 1 variable
Browsed     514 obs. of 2 variables
Cleandata   514 obs. of 3 variables
Final_data  77 obs. of 2 variables
rawdata     3087 obs. of 11 variables
Values
  badwrd    chr [1:2] "earth" "to"
  cont_an   logi [1:514] TRUE TRUE TRU…
```

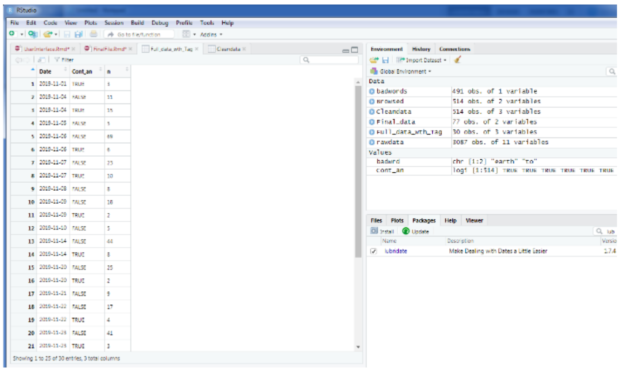
Figure 4: Dataare fetched from the MYSQL database.

The data that has been cleaned will then be filtered by the library of certain words that can be updated manually by the provider so that the words can be monitored. If the words browsed matched the library of the words, it will produce a "True" output and vice versa, at the same time, it will also show the numbers of searches made of that word, n. The filtered data will then be used to produce a graph that is easy to read and understand by users that are monitoring the browsing behaviour. Table 2 shows the words that are set for the classification method.

Table 2: Classification of words for indexing method

Class	**Words**		
Queries	Why	Tutorial	What
	Who	Which	Difference
	Whom	Does	Compare
	When	How	Meaning
Application	YouTube	Shopee	Aws
	Netflix	Lazada	Quizizz
	Spotify	Mudah	Twitch
	Flutter	Flaticon	CCNA
	Gmail	Maybank2u	Netacad
Campus Sites	iLearn	iCress	FEE
	Campus	FKE	FYP
Social Media	Facebook	Telegram	LinkedIn
	Twitter	Instagram	WhatsApp

The graph has been fetched into the Shiny Web App which is one of the R packages that provides a platform to build an interactive web app. Shiny enables the user to view the data in real-time because it provides a server that will update your data frequently**.** Figure 8 shows the output of the filtering process.



	Date	Cont_an	n
1	2019-11-01	TRUE	3
2	2019-11-04	FALSE	11
3	2019-11-04	TRUE	15
4	2019-11-05	FALSE	5
5	2019-11-06	FALSE	69
6	2019-11-06	TRUE	6
7	2019-11-07	FALSE	23
8	2019-11-07	TRUE	10
9	2019-11-08	FALSE	8
10	2019-11-09	FALSE	18
11	2019-11-09	TRUE	2
12	2019-11-10	FALSE	5
13	2019-11-14	FALSE	44
14	2019-11-14	TRUE	8
15	2019-11-20	FALSE	25
16	2019-11-20	TRUE	2
17	2019-11-21	FALSE	9
18	2019-11-22	FALSE	17
19	2019-11-22	TRUE	4
20	2019-11-23	FALSE	41
21	2019-11-23	TRUE	3

Showing 1 to 25 of 30 entries, 3 total columns

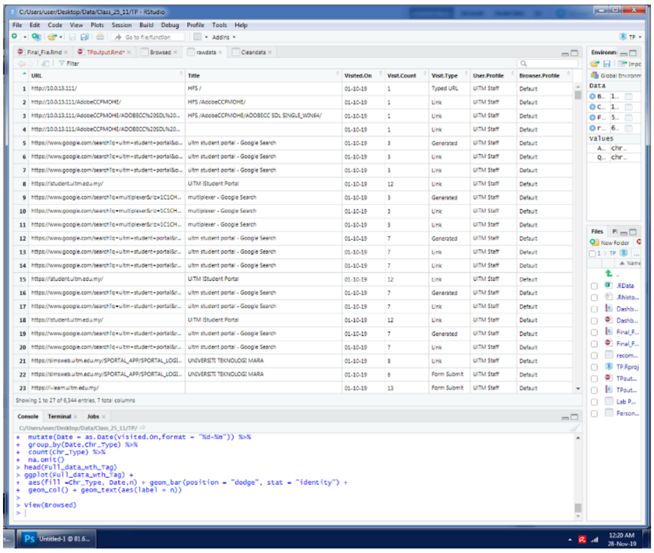
```
Data
badwords              491 obs. of 1 variable
Browsed               514 obs. of 2 variables
Cleandata             514 obs. of 3 variables
Final_data            77 obs. of 2 variables
Full_data_wth_Tag     30 obs. of 3 variables
rawdata               3087 obs. of 11 variables
Values
  badwrd              chr [1:2] "earth" "to"
  cont_an             logi [1:514] TRUE TRUE TRUE TRUE TRUE TRUE
```

	Name	Description	Versio
☑	lubridate	Make Dealing with Dates a Little Easier	1.7.4

Figure 5: Filtered data for several words browsed.

4. Result and Analysis

The data are taken from five (5) units of computers for each situation which combined and produces 20,715 browsing histories for personal computers and 6,344 browsing histories for laboratory computers. Two categories were placed inside the R script which is Applications and Queries. The raw data was fetched from the database. Figure 9 shows the raw data that has been fetched inside R programming.



	URL	Title	Visited.On	Visit.Count	Visit.Type	User.Profile	Browser.Profile
1	http://10.0.13.111/	HFS /	01-10-19	1	Typed URL	UITM Staff	Default
2	http://10.0.13.111/AdobeCCPMOHE/	HFS /AdobeCCPMOHE/	01-10-19	1	Link	UITM Staff	Default
3	http://10.0.13.111/AdobeCCPMOHE/ADOBECC%20SDL%20...	HFS /AdobeCCPMOHE/ADOBECC SDL SINGLE_WIN64/	01-10-19	1	Link	UITM Staff	Default
4	http://10.0.13.111/AdobeCCPMOHE/ADOBECC%20SDL%20...		01-10-19	1	Link	UITM Staff	Default
5	https://www.google.com/search?q=uitm+student+portal&o...	uitm student portal - Google Search	01-10-19	3	Generated	UITM Staff	Default
6	https://www.google.com/search?q=uitm+student+portal&o...	uitm student portal - Google Search	01-10-19	3	Link	UITM Staff	Default
7	https://www.google.com/search?q=uitm+student+portal&o...	uitm student portal - Google Search	01-10-19	3	Link	UITM Staff	Default
8	https://istudent.uitm.edu.my/	UiTM iStudent Portal	01-10-19	12	Link	UITM Staff	Default
9	https://www.google.com/search?q=multiplexer&rlz=1C1CH...	multiplexer - Google Search	01-10-19	3	Generated	UITM Staff	Default
10	https://www.google.com/search?q=multiplexer&rlz=1C1CH...	multiplexer - Google Search	01-10-19	3	Link	UITM Staff	Default
11	https://www.google.com/search?q=multiplexer&rlz=1C1CH...	multiplexer - Google Search	01-10-19	3	Link	UITM Staff	Default
12	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Generated	UITM Staff	Default
13	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Link	UITM Staff	Default
14	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Link	UITM Staff	Default
15	https://istudent.uitm.edu.my/	UiTM iStudent Portal	01-10-19	12	Link	UITM Staff	Default
16	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Generated	UITM Staff	Default
17	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Link	UITM Staff	Default
18	https://istudent.uitm.edu.my/	UiTM iStudent Portal	01-10-19	12	Link	UITM Staff	Default
19	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Generated	UITM Staff	Default
20	https://www.google.com/search?q=uitm+student+portal&r...	uitm student portal - Google Search	01-10-19	7	Link	UITM Staff	Default
21	https://simsweb.uitm.edu.my/SPORTAL_APP/SPORTAL_LOGI...	UNIVERSITI TEKNOLOGI MARA	01-10-19	8	Link	UITM Staff	Default
22	https://simsweb.uitm.edu.my/SPORTAL_APP/SPORTAL_LOGI...	UNIVERSITI TEKNOLOGI MARA	01-10-19	6	Form Submit	UITM Staff	Default
23	https://i-learn.uitm.edu.my/		01-10-19	13	Form Submit	UITM Staff	Default

Showing 1 to 27 of 6,344 entries, 7 total columns

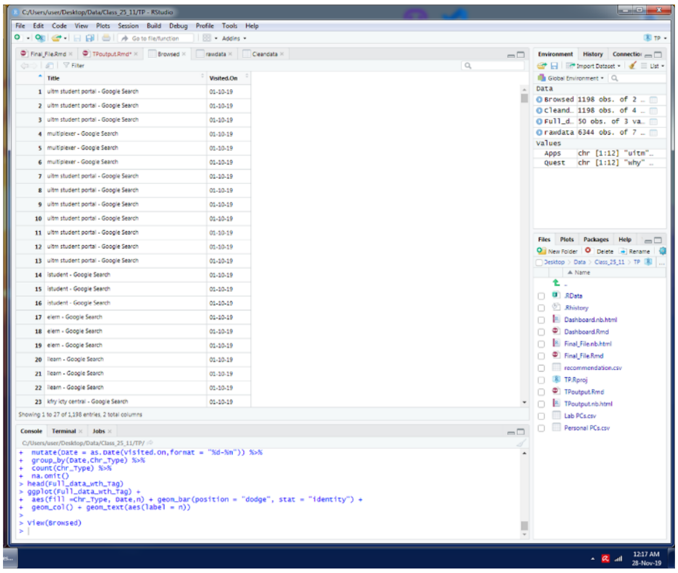
Console Terminal Jobs

C:/Users/user/Desktop/Data/Class_25_11/TP/

```
+  mutate(Date = as.Date(Visited.On,format = "%d-%m")) %>%
+  group_by(Date,Chr_Type) %>%
+  count(Chr_Type) %>%
+  na.omit()
> head(Full_data_wth_Tag)
> ggplot(Full_data_wth_Tag) +
+  aes(fill =Chr_Type, Date,n) + geom_bar(position = "dodge", stat = "identity") +
+  geom_col() + geom_text(aes(label = n))
>
> View(Browsed)
> |
```

Figure 9: Raw data was fetched into R programming.

The data cleaning process was run in which the output has taken the needed data which is the time and title for analysis. This process eliminates unnecessary data which are not the browsed words. Out of 20,715 browsing history for personal computers, only 1,999 of it was browsed words and only 1,198 out of 6,344 for laboratory computers. Analysis shows that 9.91% of data on the web browser is the real browsing behaviour for personal computers while about 18.88% is from the laboratory computers. This shows that more time is spent on learning in laboratories in the faculty than in personal space with a difference of 8.97% which is about double that of personal browsing. Figure 10 and Figure 11 show the output of cleaned data for laboratory computers and personal computers.



	Title	Visited.On
1	uitm student portal - Google Search	01-10-19
2	uitm student portal - Google Search	01-10-19
3	uitm student portal - Google Search	01-10-19
4	multiplexer - Google Search	01-10-19
5	multiplexer - Google Search	01-10-19
6	multiplexer - Google Search	01-10-19
7	uitm student portal - Google Search	01-10-19
8	uitm student portal - Google Search	01-10-19
9	uitm student portal - Google Search	01-10-19
10	uitm student portal - Google Search	01-10-19
11	uitm student portal - Google Search	01-10-19
12	uitm student portal - Google Search	01-10-19
13	uitm student portal - Google Search	01-10-19
14	istudent - Google Search	01-10-19
15	istudent - Google Search	01-10-19
16	istudent - Google Search	01-10-19
17	elern - Google Search	01-10-19
18	elern - Google Search	01-10-19
19	elern - Google Search	01-10-19
20	ilearn - Google Search	01-10-19
21	ilearn - Google Search	01-10-19
22	ilearn - Google Search	01-10-19
23	kfry icty central - Google Search	01-10-19

Showing 1 to 27 of 1,198 entries, 2 total columns

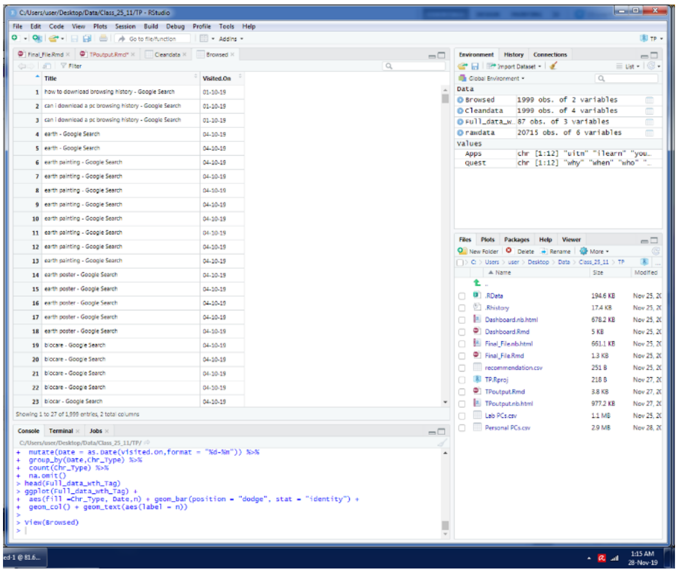
```
Data
Browsed   1198 obs. of 2 ...
Cleand... 1198 obs. of 4 ...
Full_d... 50 obs. of 3 va...
rawdata   6344 obs. of 7 ...
Values
  Apps    chr [1:12] "uitm"...
  Quest   chr [1:12] "why" ...
```

Console Terminal Jobs

C:/Users/user/Desktop/Data/Class_25_11/TP/

```
+  mutate(Date = as.Date(Visited.On,format = "%d-%m")) %>%
+  group_by(Date,Chr_Type) %>%
+  count(Chr_Type) %>%
+  na.omit()
> head(Full_data_wth_Tag)
> ggplot(Full_data_wth_Tag) +
+  aes(fill =Chr_Type, Date,n) + geom_bar(position = "dodge", stat = "identity") +
+  geom_col() + geom_text(aes(label = n))
>
> View(Browsed)
> |
```

Figure 10: The data cleaning process for browsing behaviourfor computers Laboratory.



C:/Users/user/Desktop/Data/Class_25_11/TP - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
TP
Final_File.Rmd
TPoutput.Rmd*
Cleandata
Browsed
Filter
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: > Users > user > Desktop > Data > Class_25_11 > TP
Console Terminal Jobs
ed-1 @ 81.6...
1:15 AM
28-Nov-19

	Title	Visited.On
1	how to download browsing history - Google Search	01-10-19
2	can i download a pc browsing history - Google Search	01-10-19
3	can i download a pc browsing history - Google Search	01-10-19
4	earth - Google Search	04-10-19
5	earth - Google Search	04-10-19
6	earth painting - Google Search	04-10-19
7	earth painting - Google Search	04-10-19
8	earth painting - Google Search	04-10-19
9	earth painting - Google Search	04-10-19
10	earth painting - Google Search	04-10-19
11	earth painting - Google Search	04-10-19
12	earth painting - Google Search	04-10-19
13	earth painting - Google Search	04-10-19
14	earth poster - Google Search	04-10-19
15	earth poster - Google Search	04-10-19
16	earth poster - Google Search	04-10-19
17	earth poster - Google Search	04-10-19
18	earth poster - Google Search	04-10-19
19	biocare - Google Search	04-10-19
20	biocare - Google Search	04-10-19
21	biocare - Google Search	04-10-19
22	biocare - Google Search	04-10-19
23	biocar - Google Search	04-10-19

Showing 1 to 27 of 1,999 entries, 2 total columns

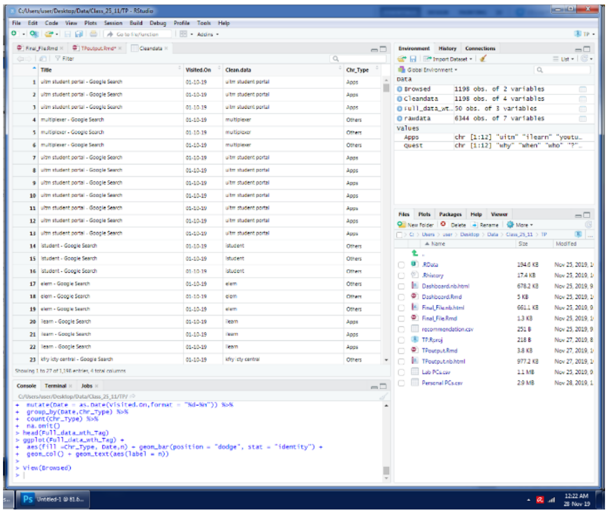
```
Data
Browsed        1999 obs. of 2 variables
Cleandata      1999 obs. of 4 variables
Full_data_w... 87 obs. of 3 variables
rawdata        20715 obs. of 6 variables
Values
  Apps         chr [1:12] "uitm" "ilearn" "you...
  quest        chr [1:12] "why" "when" "who" "...
```

Name	Size	Modified
..		
.RData	194.6 KB	Nov 25, 20
.Rhistory	17.4 KB	Nov 25, 20
Dashboard.nb.html	678.2 KB	Nov 25, 20
Dashboard.Rmd	5 KB	Nov 25, 20
Final_File.nb.html	661.1 KB	Nov 25, 20
Final_File.Rmd	1.3 KB	Nov 25, 20
recommendation.csv	251 B	Nov 25, 20
TP.Rproj	218 B	Nov 27, 20
TPoutput.Rmd	3.8 KB	Nov 27, 20
TPoutput.nb.html	977.2 KB	Nov 27, 20
Lab PCs.csv	1.1 MB	Nov 25, 20
Personal PCs.csv	2.9 MB	Nov 28, 20

```
C:/Users/user/Desktop/Data/Class_25_11/TP/
+  mutate(Date = as.Date(visited.on,format = "%d-%m")) %>%
+  group_by(Date,Chr_Type) %>%
+  count(Chr_Type) %>%
+  na.omit()
> head(Full_data_wth_Tag)
> ggplot(Full_data_wth_Tag) +
+  aes(fill =Chr_Type, Date,n) + geom_bar(position = "dodge", stat = "identity") +
+  geom_col() + geom_text(aes(label = n))
>
> View(Browsed)
> |
```

Figure 11: The data cleaning process for browsing behaviour for Personal computers.

Later, the cleaned data were classified at the classification process. The classifications were based on the library made which is applications or queries. Data checking was done, if the data is not in either of the libraries, it falls into the other group. This is where the process of creating insight was done. The bigger the library of the classification, the accurateness of the analytics increases. This is a repetitive process where the words will try and error until it finds the match words and placed them inside their group. This process used the Lubridate Package in the R programming library to group them by dates. The output of this process is the data that is used to plot the graph of insight which consists of the application, queries, social media, Campus related sites. Data which is are not fit to the groups are places in other groups. Figure 12 and Figure 13 show the output of the classification of data for the computer laboratory data and personal computer data.



C:/Users/user/Desktop/Data/Class_25_11/TP - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
TP
Final_File.Rmd
TPoutput.Rmd*
Cleandata
Filter
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: > Users > user > Desktop > Data > Class_25_11 > TP
Console Terminal Jobs
Untitled-1 @ 81.6...
12:22 AM
28 Nov 19

	Title	Visited.On	Clean.data	Chr_Type
1	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
2	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
3	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
4	multiplexer - Google Search	01-10-19	multiplexer	Others
5	multiplexer - Google Search	01-10-19	multiplexer	Others
6	multiplexer - Google Search	01-10-19	multiplexer	Others
7	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
8	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
9	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
10	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
11	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
12	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
13	uitm student portal - Google Search	01-10-19	uitm student portal	Apps
14	istudent - Google Search	01-10-19	istudent	Others
15	istudent - Google Search	01-10-19	istudent	Others
16	istudent - Google Search	01-10-19	istudent	Others
17	elem - Google Search	01-10-19	elem	Others
18	elem - Google Search	01-10-19	elem	Others
19	elem - Google Search	01-10-19	elem	Others
20	ilearn - Google Search	01-10-19	ilearn	Apps
21	ilearn - Google Search	01-10-19	ilearn	Apps
22	ilearn - Google Search	01-10-19	ilearn	Apps
23	kfry icty central - Google Search	01-10-19	kfry icty central	Others

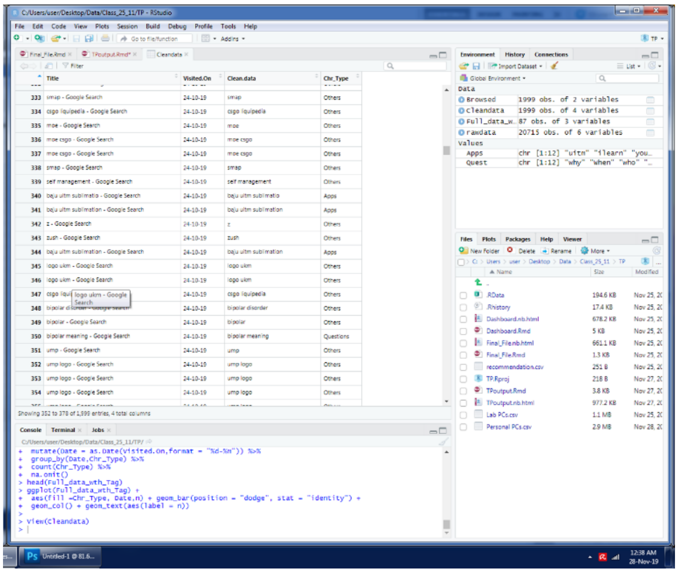
Showing 1 to 27 of 1,198 entries, 4 total columns

```
Data
Browsed        1198 obs. of 2 variables
Cleandata      1198 obs. of 4 variables
Full_data_wt... 50 obs. of 3 variables
rawdata        6344 obs. of 7 variables
Values
  Apps         chr [1:12] "uitm" "ilearn" "youtu...
  quest        chr [1:12] "why" "when" "who" "?"...
```

Name	Size	Modified
..		
.RData	194.6 KB	Nov 25, 2019, 1
.Rhistory	17.4 KB	Nov 25, 2019, 1
Dashboard.nb.html	678.2 KB	Nov 25, 2019, 9
Dashboard.Rmd	5 KB	Nov 25, 2019, 1
Final_File.nb.html	661.1 KB	Nov 25, 2019, 9
Final_File.Rmd	1.3 KB	Nov 25, 2019, 1
recommendation.csv	251 B	Nov 25, 2019, 9
TP.Rproj	218 B	Nov 27, 2019, 8
TPoutput.Rmd	3.8 KB	Nov 27, 2019, 1
TPoutput.nb.html	977.2 KB	Nov 27, 2019, 1
Lab PCs.csv	1.1 MB	Nov 25, 2019, 9
Personal PCs.csv	2.9 MB	Nov 28, 2019, 1

```
C:/Users/user/Desktop/Data/Class_25_11/TP/
+  mutate(Date = as.Date(visited.on,format = "%d-%m")) %>%
+  group_by(Date,Chr_Type) %>%
+  count(Chr_Type) %>%
+  na.omit()
> head(Full_data_wth_Tag)
> ggplot(Full_data_wth_Tag) +
+  aes(fill =Chr_Type, Date,n) + geom_bar(position = "dodge", stat = "identity") +
+  geom_col() + geom_text(aes(label = n))
>
> View(Browsed)
> |
```

Figure 12: Classified data for computer laboratory browsing behaviour.



C:/Users/user/Desktop/Data/Class_25_11/TP - RStudio

	Title	Visited.On	Clean.data	Chr_Type
333	smap - Google Search	24-10-19	smap	Others
334	csgo liquipedia - Google Search	24-10-19	csgo liquipedia	Others
335	moe - Google Search	24-10-19	moe	Others
336	moe csgo - Google Search	24-10-19	moe csgo	Others
337	moe csgo - Google Search	24-10-19	moe csgo	Others
338	smap - Google Search	24-10-19	smap	Others
339	self management - Google Search	24-10-19	self management	Others
340	baju uitm sublimatio - Google Search	24-10-19	baju uitm sublimatio	Apps
341	baju uitm sublimation - Google Search	24-10-19	baju uitm sublimation	Apps
342	z - Google Search	24-10-19	z	Others
343	zush - Google Search	24-10-19	zush	Others
344	baju uitm sublimation - Google Search	24-10-19	baju uitm sublimation	Apps
345	logo ukm - Google Search	24-10-19	logo ukm	Others
346	logo ukm - Google Search	24-10-19	logo ukm	Others
347	csgo liqui logo ukm - Google Search	24-10-19	csgo liquipedia	Others
348	bipolar disorder - Google Search	24-10-19	bipolar disorder	Others
349	bipolar - Google Search	24-10-19	bipolar	Others
350	bipolar meaning - Google Search	24-10-19	bipolar meaning	Questions
351	ump - Google Search	24-10-19	ump	Others
352	ump logo - Google Search	24-10-19	ump logo	Others
353	ump logo - Google Search	24-10-19	ump logo	Others
354	ump logo - Google Search	24-10-19	ump logo	Others

Showing 352 to 378 of 1,999 entries, 4 total columns

C:/Users/user/Desktop/Data/Class_25_11/TP/

```
+  mutate(Date = as.Date(visited.On,format = "%d-%m")) %>%
+  group_by(Date,Chr_Type) %>%
+  count(Chr_Type) %>%
+  na.omit()
> head(Full_data_wth_Tag)
> ggplot(Full_data_wth_Tag) +
+  aes(fill =Chr_Type, Date,n) + geom_bar(position = "dodge", stat = "identity") +
+  geom_col() + geom_text(aes(label = n))
>
> View(Cleandata)
> |
```

```
Data
Browsed        1999 obs. of 2 variables
Cleandata      1999 obs. of 4 variables
Full_data_w…   87 obs. of 3 variables
rawdata        20715 obs. of 6 variables
Values
Apps           chr [1:12] "uitm" "ilearn" "you…
Quest          chr [1:12] "why" "when" "who" "…
```

C: > Users > user > Desktop > Data > Class_25_11 > TP

Name	Size	Modified
..		
.RData	194.6 KB	Nov 25, 20
.Rhistory	17.4 KB	Nov 25, 20
Dashboard.nb.html	678.2 KB	Nov 25, 20
Dashboard.Rmd	5 KB	Nov 25, 20
Final_File.nb.html	661.1 KB	Nov 25, 20
Final_File.Rmd	1.3 KB	Nov 25, 20
recommendation.csv	251 B	Nov 25, 20
TP.Rproj	218 B	Nov 27, 20
TPoutput.Rmd	3.8 KB	Nov 27, 20
TPoutput.nb.html	977.2 KB	Nov 27, 20
Lab PCs.csv	1.1 MB	Nov 25, 20
Personal PCs.csv	2.9 MB	Nov 28, 20

12:38 AM
28-Nov-19

Figure 13: Classified data for personal computer browsing behaviour.

The data of classified browsing is then imported into a graph of browsing date versus the frequency of browsing. The analysis presents all the combined classified browsing words to show the trend of browsing of each situation. Figure 14 shows the analysis of browsing behaviour for laboratory computers. It is identified that other criteria are the highest that shows 72.4% browsing compared to 0% has browsed on social media, 8% for campus site pages, 12.2% for browsing queries, and 7.4% browsed for applications.

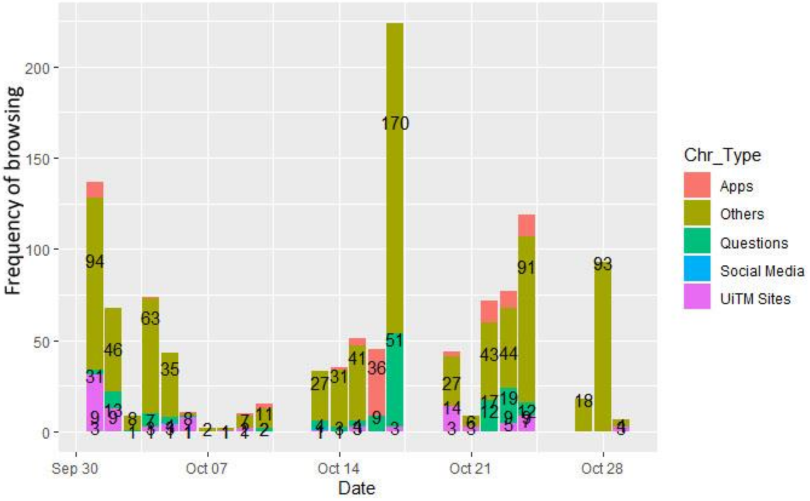


Figure 14: Browsing behaviour of laboratory computers.

Figure 15 shows the analysis of browsing behaviour for personal computers. It is identified that other criteria are the highest that shows 76.9% browsing compared to 0.3 browsed on social media, 5.3% for campus site pages, 10.1% for browsing queries, and 7.7% browsed for applications.

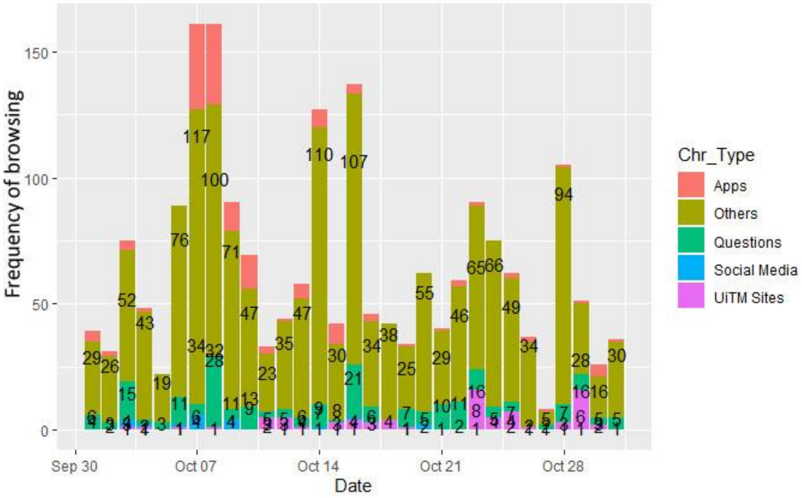


Figure 15: Browsing behavior of personal computers.

Based on browsing behaviour, it is identified that the other category scored the highest, thus more libraries should be filtered for the analysis on the data mining programmed. The R script will then group the data into types of browsing into the monthly record. Figure 16 and Figure 17 show the group of browsing in a month.

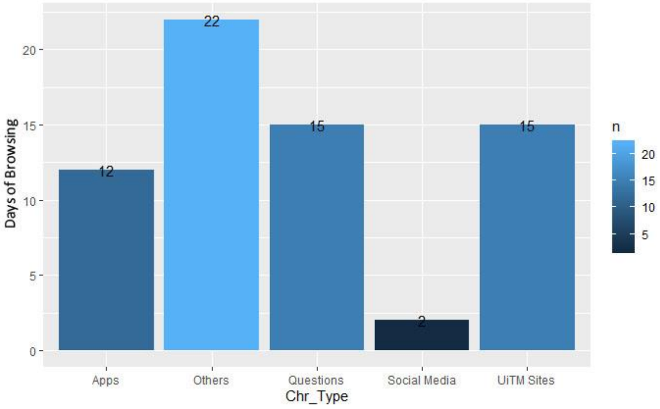


Figure 16: Types of browsing laboratory computers.

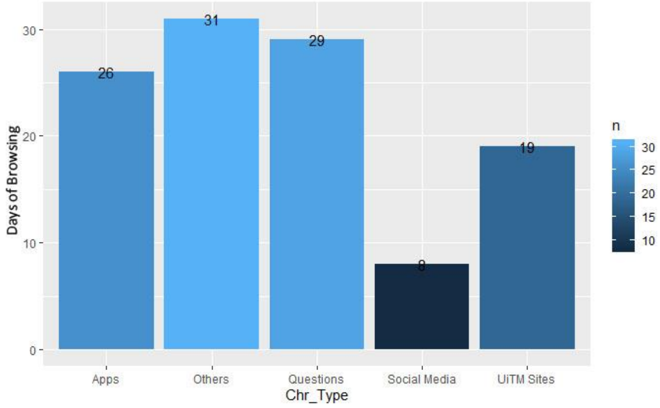


Figure 17: Types of browsing of personal computers.

Finally, all the processed data are imported into a dashboard to ease the user for monitoring purposes. All the data needed can be attached to the dashboard using the Shiny package in R programming. As the data for this research increase, it can then generate an equation that can determine patterns and predicts future outcomes and trends. Supervised learning is the process of creating predictive models using a set of historical data that contains the results you are trying to predict(Shmueli & Koppius, 2011). Parents or employers can now monitor unethical browsing just by adding the words in the library. Figure 18 shows the dashboard of this research.

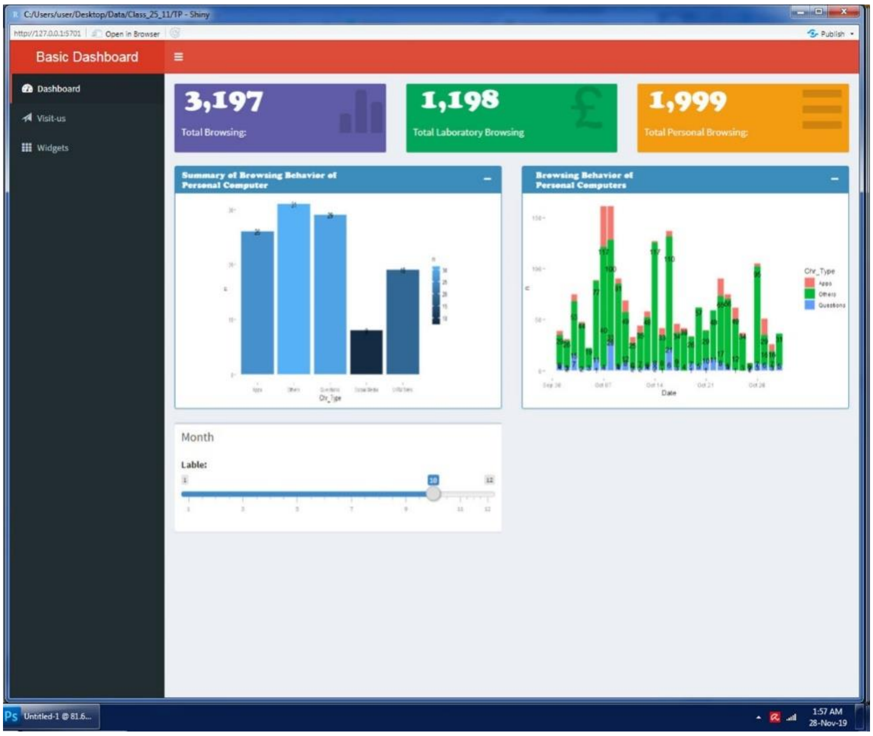


Figure 18: The dashboard that shows all the result of this research using Shiny package.

5. Conclusion

Web browsing behaviours have been studied in many aspects from web usage mining, applications, images downloadable, personalized systems, and many more. Web browsing has exposed students and more people use the Web to communicate, work, and have fun. Thus, monitoring or analysing is possible to identify someone based on their Web browsing behaviours or to differentiate between two persons based solely on their Web browsing histories. Based on this study, this paper provides some insights into students' web browsing behaviours using data mining at a campus network. We describe characteristic features of Web browsing behaviours and present our analysis for those features for user authentication based on browsed words in categories. Categorizing words that were browsed by the students in two browsing situations which are personal computers and laboratory. The categories in groups for social media, queries, applications, and Campus related sites, and others which is not in the categories group. The result of the research shows that the amount of browsing using personal computers is way higher than the laboratory due to the accessibility of the computer. However, using this data mining categories group shows that others that are not in the four (4) groups are much higher. Thus, future study to look in detail of browsing behaviours sites is to be done in more detailed categories. This analysisresults will help universities to monitor the browsing behaviour of their students' surrounding especially the usage of the internet in the computers' lab.

Acknowledgment

The authors would like to thank the Research Management Institute, Universiti Teknologi MARA for the support grant in publishing this paper.

References

1. Abramson, M., & Aha, D. W. (2013). *User authentication from web browsing behavior*. https://apps.dtic.mil/sti/citations/ADA599778
2. Adadi, A. (2021). A survey on data-efficient algorithms in big data era [Article]. *Journal of Big Data*, *8*(1), Article 24.doi:10.1186/s40537-021-00419-9
3. Astika Saputra, F., Barakbah, A., & Riza Rokhmawati, P. (2020). Data Analytics of Human Development Index(HDI) with Features Descriptive and Predictive Mining. IES 2020 - International Electronics Symposium: The Role of Autonomous and Intelligent Systems for Human Life and Comfort, doi:10.1109/IES50839.2020.9231661
4. Bursztein, G. A. E., Jackson, C., & Boneh, D. (2010). An analysis of private browsing modes in modern browsers. Proceedings of the 19th USENIX Security Symposium, doi:10.5555/1929820.1929828
5. Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, *115*(9).doi:10.1.1.695.5829
6. Diep, N. N., Van Tien, N., Anh, N. H., & Phuong, T. M. (2019). An unsupervised method for web user interest analysis. Proceedings - 2019 6th NAFOSTED Conference on Information and Computer Science, NICS 2019, doi:10.1109/NICS48868.2019.9023842
7. Gao, S., Liu, Y., Kang, Y., & Zhang, F. (2021). User-Generated Content: A Promising Data Source for Urban Informatics. 503-522 doi: 10.1007/978-981-15-8983-6_28
8. Jalil, N. A., Hwang, H. J., & Dawi, N. M. (2019). Machines learning trends, perspectives and prospects in education sector. ACM International Conference Proceeding Series, doi:10.1145/3345120.3345147
9. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects [Review]. *Science*, *349*(6245), 255-260.doi:10.1126/science.aaa8415
10. Kassim, M., Ismail, N. L., Mohamad, R., Suliman, S. I., & Ismail, M. (2017). Self-similarity hurst parameter estimation with rescaled range method on ip-based campus internet traffic [Article]. *Pertanika Journal of Science and Technology*, *25*(S4), 287-302. http://mymedr.afpm.org.my/publications/57671
11. Kaushik, A., & Jones, G. (2018). *Exploring Current User Web Search Behaviours in Analysis Tasks to be Supported in Conversational Search*.doi:arxiv.org/abs/2104.04501
12. Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques [Review]. *Informatica (Ljubljana)*, *31*(3), 249-268.doi:10.5555/1566770.1566773
13. Labaj, M., & Bieliková, M. (2015). Conducting a web browsing behaviour study – an educational scenario. 8939. 531-542 doi: 10.1007/978-3-662-46078-8_44
14. Lee, W., Lee, S. S., Chung, S., & An, D. (2007). Harmful contents classification using the harmful word filtering and SVM. International Conference on Computational Science, doi:10.1007/978-3-540-72588-6_3
15. Liu, S., & Liu, H. (2021). Tagging Items Automatically Based on Both Content Information and Browsing Behaviors. *INFORMS Journal on Computing*.doi:10.1287/ijoc.2020.1007
16. Mathur, S., Nikam, P., Patidar, H., Gaikwad, R. B., & Nayak, P. N. (2021). Machine-Learning directed Article Detection on the Web using DOM and text-based features. 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), doi:10.1109/CCNC49032.2021.9369599
17. Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making [Article]. *Big data*, *1*(1), 51-59. doi:10.1089/big.2013.1508
18. Sharma, S., Singla, K., Rathee, G., & Saini, H. (2020). A hybrid cryptographic technique for file storage mechanism over cloud. 1045. 241-256 doi: 10.1007/978-981-15-0029-9_19
19. Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research [Review]. *MIS Quarterly: Management Information Systems*, *35*(3), 553-572.doi: 10.2307/23042796
20. Surya Gunawan, T., Aleah Jehan Abdullah, N., Kartiwi, M., & Ihsanto, E. (2020). Social Network Analysis using Python Data Mining. 2020 8th International Conference on Cyber and IT Service Management, CITSM 2020, doi:10.1109/CITSM50537.2020.9268866
21. Toba, H., Jomei, C. S., Setiawan, L., Karnalim, O., & Il, H. (2020). Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters. 2020 8th International Conference on Information and Communication Technology, ICoICT 2020, doi:10.1109/ICoICT49345.2020.9166179
22. Zamfiroiu, A., & Sbora, C. (2014). Statistical analysis of the behavior for mobile E-learning. *Procedia Economics and Finance*, *10*, 237-243. doi: 10.1016/S2212-5671(14)00298-6