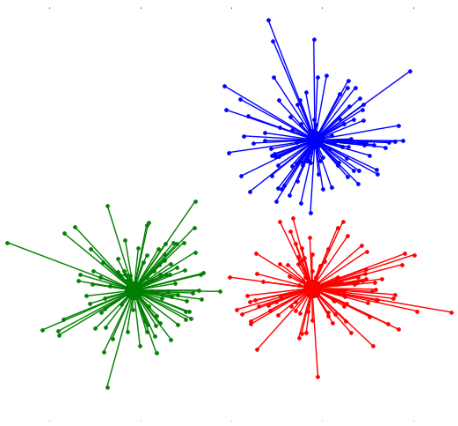

Figure 4 Example for Three Clusters

In the web usage area, there are two types of clusters frequently used: usage clusters and page clusters [54]. Clustering of users aims to create groups of users having similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-commerce applications or provide personalised web content to the users. On the other hand, clustering of pages will discover groups of pages that have related content. This information is useful for Internet search engines and web assistance providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs. Clustering techniques can be used to recognise the natural groupings of customers and find an answer founded on observed data patterns. Providing the data mining models are properly built; they can discover groups with diverse profiles and features and guide to a lot of segmentation patterns with business insight and vision [55]. These segments are essential in order to give precedence to customer handling and marketing interventions according to the significance of each customer.

K-means:

Many different clustering algorithms exist, but one of the most commonly used is K-means clustering proposes to separate observations into k clusters in which each observation is in the right lace to the cluster with the nearest mean, operating as a prototype of the cluster [56]. It is a distance-based clustering technique and it does not need to compute the distances between all combines of records. The amount of clusters to be designed is prearranged and determined by the user in advance. Typically an amount of several solutions should be tested before accepting the most appropriate. It is best for handling continuous clustering fields [55].

c) *Big Data*

Nowadays, a huge amount of data is being collected and stored from web data, e-commerce, bank and social network. Big data is the term use for large-volume and complex data, that it becomes difficult to process it using typically database management systems or data processing applications. The challenges consist of the areas of collection, storage, search, sharing, transfer, analysis, and visualization of this data [57].

In e-business area, big data refers to the huge quantities of transaction, click-stream, voice, and video data in the e-commerce landscape. In general, e-business websites have characteristics with both structured and unstructured data [58]. Structured data belongs to kinds of data with a high level of organization, such as information in a relational database such as age, gender, date of birth, address, and preferences. Whereas unstructured data refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is normally text-heavy, but may hold data such as dates, numbers, and facts as well.

Big data can be explained in terms of five Vs: volume, velocity, variety, veracity, and value [59]. The ‘volume’ refers to the quantities of big data, which is increasing exponentially. The ‘velocity’ is the speed of data collection, processing and analyzing in the real time. The ‘variety’ refers to the different types of data collected in big data environment. The ‘veracity’ represents the reliability of data sources. the ‘value’ represents the transactional, strategic and informational benefits of big data. Figure 5 shows the main characterization of Big Data.

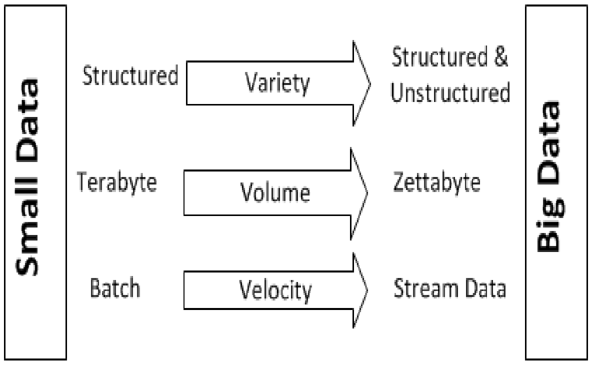

Figure 5 Main characterization of Big Data

D. Web Mining

Web usage mining is a systematic way of using data mining techniques to find usage patterns from web data, including web documents, hyperlinks between documents and usage logs of websites [60]. Web mining can help to understand customer behaviour and evaluate the performance of a website.

Web mining can be classified into three different varieties as shown in Figure 6. These are: web usage mining, web content mining and web structure mining [61].

Web mining consists of three phases [62]:

- Discovering resources
- Choosing information and pre-processing
- Extracting knowledge and analyzing patterns.

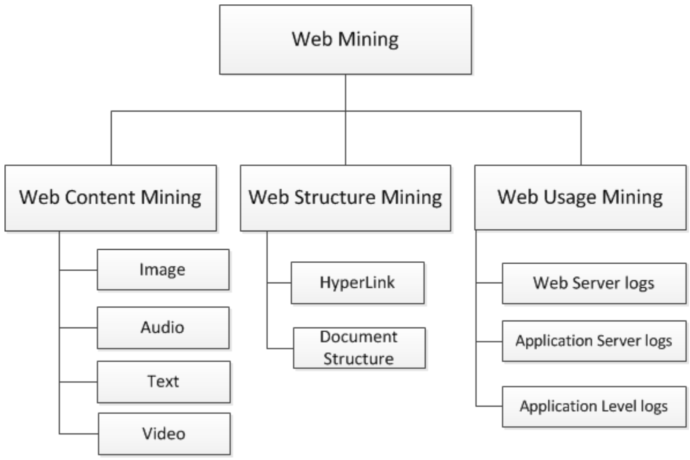

Figure 6 Web mining classification

As shown in Figure 6, the data obtained through different sources can be classified into three main groups, namely: web usage data, web content data and web structure data. Table 3 shows web mining techniques, objective, collection source and examples of their applications.

		Objective	Collection source	Applications
Web mining techniques	Web Content	Index:extraction knowledge from Web page contents	Text pages	• Text Mining • Opinion Mining • Website Improvement
	Web Structure	Map:discover useful knowledge from the hyperlinks structure	Hyperlinks	• Web Page Rating • Web Clustering • Web Classification

	Web Usage Mining	Behaviour:discover user access patterns	User accessing, Logs	• Navigational Patterns • Session and Visitor Analysis • Business Intelligent System

Table 3 Web mining techniques, objective, collection source and example of their applications.

a) *Web Usage Mining*

Web usage mining is the application of data mining techniques to discover patterns from the websites [60]. The log data is collected automatically by the web and application servers act for the fine-grained navigational behaviour of visitors , It is considered the main source of data in web usage mining [63]. The operational database(s) for the site may include additional user profile information. The data may include demographic information about registered users, user ratings on various objects such as products or movies, past purchases or visit histories of users, as well as other explicit or implicit representations of users' interests.

However, web usage mining aims to find an interesting knowledge from the web data gained from the interactions of the users with the web. Web usage mining uses data mining techniques to analyse search or other activity logs to find interesting patterns. The purpose of web usage mining is to recognise the behaviour of Internet users through the process of data mining techniques. Knowledge gained from web usage mining can be applied to improve web design, present a personalised service and facilitate more effective browsing. One of the main applications of web usage mining is to create customer profiles. Web usage mining has become essential for operational web site management. For example, the quality of services could be improved by applying web usage mining as companies can recognise the needs of their customers and therefore take necessary action to respond to their needs. In addition, companies can identify, attract and keep customers [64].

Web usage mining is used to discover usage patterns from web data, in order to recognise and improve the needs of websites such as personalised services, adaptive web sites, customer profiling and creating attractive web sites. This is can be done by applying data mining techniques to discover user's pattern. These patterns are used to understand the core features of the users' behaviours in order to improve the website structure and establish personal or dynamic recommendations about content of the web [65]. For example, applying data mining techniques on web access logs can help to identify the user behaviour and the web structure. From the business perspective, knowledge gained from the web usage patterns could be used to usefully manage activities related to e-business as showed in Figure 7.

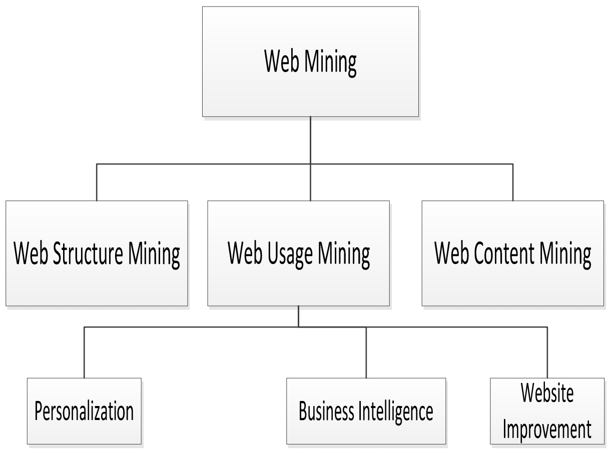

Figure 7 Web Usage Mining in E-Business

The usage patterns extracted from web data can be used in a different range of web applications such as web personalisation, analysing page sequences, website improvement, site modification, and business intelligence discovery usage characterisation.

Data type	Description	Data Examples
Web Server Data	The user logs are collected via HTTP.	IP address, page reference and access time.
Application Server Data	Contains data generated by web application server. This application server hosts business layer	Users transactions and encryption data
Application Level Data	Data are collected from an application.	Trace messages, Debug messages, Information messages, Warning messages, Error messages.

Table 4 Web usage mining sources

Web server data, application data server, and application level data can very easily collect data about web usage, as shown in Table 4. There are various resources for web usage mining as shown in Table 4, such as web access logs, cookies, data tags, login information, client or server side scripts, packet sniffing. However, web access logs are the major sources for web usage mining. They are recorded in standardised text file format used by web servers when generating server log files [64]. Because the format is standardised, the files can be readily analysed by a variety of web analysis software packages. Web server logs store data about every visit to the website hosted on a server. For example, when a web user visited a web page, valuable information can be stored on web server data such as the Internet Protocol (IP) address of the request, the error code, and the number of bytes sent to the user, and the type of browser used. Logs file can also be stored by web server, which shows the page from which a web user makes the next request. Client-applications, such as web user's browsers can also be utilised to store a user's actions.

Customer's behaviour and useful information can be discovered by application data mining and performing analysis on websites. This knowledge has numerous applications, such as personalisation and collaboration in web-based systems, marketing, website design, website evaluation, and decision support. Web logs usually consist of usage data for more than one user. Web usage mining can help identify users who have accessed similar web pages. The patterns that emerge can be applied in collaborative web searching and collaborative filtering.

However, the main objective of web usage mining is to find interesting information about customers' patterns. One of the main difficulties confronted by web usage mining applications is that web server log data are unidentified, creating a challenge to recognise users and user sessions from the user transaction. Techniques like web cookies and user registration have been applied in some applications, but each technique has its weaknesses. In pattern discovery, data mining techniques, such as association rule mining, classification and clustering, can be used. For example, Munk et al. (2010) applied clustering on web log data to recognise users who have browsed similar web pages.

Authors	Propose Model
Perkowitz and Etzioni (2000)	They proposed the idea of optimising the structure of web sites based on co-occurrence patterns of pages within usage data for the site [66].
Schechter et al (1998)	They developed techniques for using path profiles of users to predict future HTTP requests, which can be used for network and proxy caching [67] .
Castellano et al (2013)	Have applied data mining techniques to extract usage patterns from web logs, for the

	purpose of deriving marketing intelligence [68].
Eirinaki and Vazirgiannis (2003)	Have proposed clustering of user sessions to predict future user behaviour [69] .
Yadav (2012)	Clustering the buyer's behaviour on ecommerce site depending on their age [29]
Carmona et al.(2012a)	Applied data mining techniques used in an e-commerce website of extra virgin olive oil to provide some guidelines for improving its usability and user satisfaction [70].
Verma (2015)	Used semantic web mining and e-neural computing to improve the page ranking and help the website designers to optimise the website structure [47] .

Table 5 Different approaches applying web usage mining in websites.

Table 5 shows the web usage mining has been used to evaluate the websites which helped to optimise the website structure and improve user satisfaction.

III. METHODOLOGY

In this research, several features related mainly with the user behaviours in purchasing are analyzed and then rank those features to understand which user behaviours feature maximize conversion in order to discover new knowledge and extract a new business insight. This will help to increase the conversion rate and help the business to grow. Also, classify and predict the buyers and non-buyers. In order to do this, the following approach will be used: this approach contains 7 stages starting from understanding the business requirements to ranking the features. A high-level overview of the methodology is presented in Figure 2.

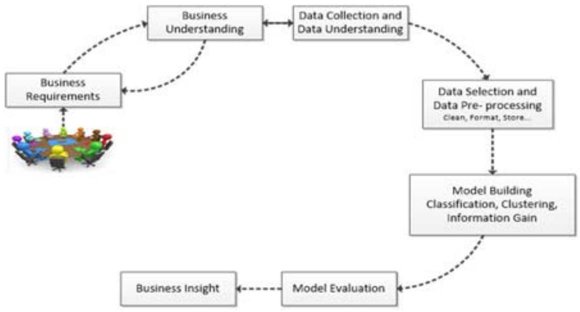

Figure 2 Ranking algorithm approach

A. *Business Requirements and Business Understanding*

This primary stage focuses on understanding the business objectives and business requirements for this research from a business view. A business object is analysing the websites that are suffering from low conversion rates to understand which user's behaviours are responsible for this poor conversion. Next, our research goal is to increase the conversion rates for the websites. The business goal translated into a data mining objective and development of a project plan. The business purpose is very important for the model building. For example, understanding how sales revenue is related to website evaluation to optimise the website structure. Therefore, the business objective should be interpreted as a data mining objective.

B. *Data Collection and Data Understanding*

The data understanding stage begins with a preliminary data collection of necessary information, proceeds with events in order to get aware with the data and to find first insights into the data. This data lets companies bring into line their websites' aims with their business objectives for the purpose of recognising areas for enhancement, promoting popular parts of the site and eventually increasing revenue.

Data requirements are considered, understanding the meaning of each feature and what the purpose is for this feature as well as what knowledge we can discover from this feature. In addition, understanding how web analytics tools collect data, process data and generate reports. For example, engagement, which holds a visitor's attention, presents how long a visitor can stay on a specific website. However, if a visitor only stays any page and quits then analytics tools sets them in the zero to 10 seconds classification.

The data will be collected using webmaster tools such as Google Search Console API and Google Analytics [71]. The data is collected from different websites, Web analytics such as webmaster tools are a free service that gives statistics and essential analytical tools for search engine optimization (SEO) and marketing purposes. This service is provided by Google. Google Analytics is currently the most commonly used web analytics service on the Internet. Developers can integrate the data from web analytics tools into existing products or create standalone applications that can be built on several processes that have been applied in a data set in order to import the visitor's data. Table 6 shows the List of the user's behaviour features.

Feature name	**Description**

Browser	This feature contains the browser which is used by the user when visiting. Such as Internet Explorer, Firefox, Chrome, Safari or a mobile browser.
City	The cities of users, derived from IP addresses or geographical IDs.
Source	This feature describes the source used by the user to access the website. Direct (D): Access performed directly to the website address. Engine (E): Access performed through a search engine.
OS	The operating system is low-level software that handles computer hardware and software assets and delivers basic public functions for computer programs.
User ID	A logical entity to identify the user on a website or within any generic IT environment. It is used to distinguish between the users who access the website.
Sessions	A session is a time of action by an individual web browser from the arrival point to departure point.
View Duration	The duration of user sessions represented in total seconds.
Page View	The total number of page views on the website.
Transaction	The overall of browsers who buy services or goods.
Transaction Revenue	The total sale revenue provided in the transaction excluding shipping and tax (total income cash or credit for goods, services or assets).
Transaction Quantity	The total number of items purchased. For example, two items have been purchased.
Website ID	A logical entity to identify the website within any generic IT environment. It is used to distinguish between the websites.

Table 6 List of the user's behaviour features

C. Data Selection and Data Pre-Processing

This stage is about the process of retrieving data from the websites by using webmaster tools, the websites provide a huge amount of data but unfortunately, this data is usually unstructured and requires a long process. There is a need to develop the APIs which extract the relevant data from the various third-party tools (such as Google AdWords and Google Analytics) and present this in meaningful ways to business owners and marketers, meaning current and future predicted website issues can be addressed.

Since E-business websites are high-volume, high-velocity and high-variety information. Therefore, the effective and innovative forms of information processing are necessary to assist enhance insight, decision making. Figure 3 shows the process that I have used to transfer the raw data which is extracted from webmaster tools in order to prepare it for the data mining stage. Tasks involve table, record and feature selection, in addition to conversion and cleaning of data for building the model.

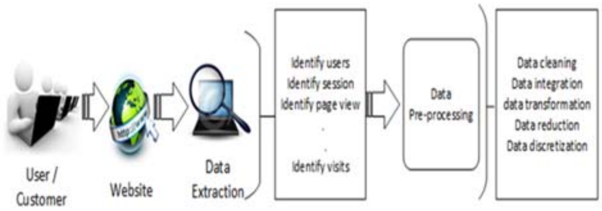

Figure 3 Data Pre-processing stage

Since the data has different issues such as redundancy and unstructured data, opacity and incomplete, pre-processing this data is essential in order to solve these issues and become ready for applying data mining techniques. For these purposes, the database management system, specifically MYSQL, has been used to manipulate the data in order to clean the data then applying data mining techniques. Row data has converted to popular Comma Separated Values (CSV) format.

This stage includes the acquisition, integration, and formatting of the data corresponding to the data mining requirements. The consolidated data have to be ‘‘clean’’ and correctly converted according to the requirements of the specific data mining techniques that would be applied. After the data is converted into an understandable format data mining techniques can be applied. This process is critical to the successful extraction of the website visitor pattern. It is a process that involves several tasks and which cannot be totally automated. The process requires pre-processing of the primary data, combining information from various websites, and converting the combined machine learning techniques. This is mainly significant in websites because of a series of mouse clicks made by a customer while accessing the e-commerce websites and its association with other related data collected from several sources. Usage data preparation brings a sum of difficulties, which requires the use of a range of algorithms and database management techniques for pre-processing tasks. However, inadequate pre-processing of data usually results in inaccurate and unreliable data mining results.

The final set of attributes consists of User_ID, Sessions, ViewDuration, Pageviews, Transactions, Browser, Operating System, City, transaction revenue and ItemQuantity, Website_ID.

D. Model Building

In this stage, modeling techniques are selected. In particular, the building and evaluation of alternative modeling algorithms, dividing of the dataset into training and testing subgroups for assessment purposes. The handled data are then applied for model training. Analysts should select the suitable modeling methods for the specific business purpose. In advance, the training of the models and especially in the case of predictive modeling, the modeling dataset should be divided so that the model's performance is assessed on an isolated dataset. A specific mathematical algorithm is applied to the pre-processing data in order to extract website issues. This also involves ingesting and wrangling of a wide variety of relevant data sources to profile, monitor and measure website performance to predict future performance issues, using a combination of supervised and unsupervised learning approaches.

E. Model Evaluation

The conversion rate is the percentage of customers who browse a commercial website and make a purchase. It is very important to know when the customer who comes to the website and “convert” to the buyer because the conversion rate can be calculated. Next, Conversion rate can be used to predict future success or use it to determine that something isn’t working. In addition, conversion rates are actually useful to website owners, who can use website traffic data to figure out what other marketing processes should be used to raise product sales. Moreover, we can track the online user behaviour previous to buying and finding interesting information about the buyer and non-buyer such as demographic analysis of who purchases, and the browser platform that users are mainly using. This information is vital to increase leads and revenue for any website.

The created models will then be evaluated not only in terms of technical perspective but also, more significantly, in terms of the business success conditions set out in the business understanding phase. In this stage the best model is chosen that represents the websites evaluation issues and how well the chosen model will work. A specific table layout will use that allows visualization of the performance of an algorithm called Confusion Matrix. The table reports the number of false positives, false negatives, true positives, and true negatives and provides analysis details that correct guesses (accuracy).

Our model is evaluated based on the following standard performance measures:

True positive (TP): Number of correctly predicted as a visitor conversion.

True negative (TN): Number of correctly predicted as a non-visitor conversion.

False positive (FP): Number of wrongly predicted a visitor conversion, when a detector predicted buyer person is a not a buyer.

False negative (FN): Number of wrongly predicted as non-visitor conversion, when a detector fails to identify the non-visitor conversion.

Recall is the percentage of correct positives that are truly predicted by the classifier.

F-measure is defined as the harmonic mean of recall and precision according to the following equation:

Table 7 shows the confusion matrix for a two-class classifier. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

	Predicted buyer	Predicted non-buyer
Actual buyer	True positive (TP)	False positive (FP)
Actual non-buyer	False negative (FN)	True negative (TN)

Table 7 Confusion Matrix

Table 8 shows a sample data set used for classifying customers into buyers or non-buyers.

Users	new users	Sessions	Session	Page	bounces	Browser	OS	City	Class
2	2	2	522	13	0	Firefox	Windows	Brisbane	No
0	0	1	2476	5	0	Explorer	Windows	Brisbane	Yes
37	37	37	51	40	34	Chrome	Windows	Wollongong	No
1	1	2	11	3	1	Chrome	Macintosh	Auckland	Yes
21	21	25	7452	159	5	Chrome	Windows	Auckland	NO
1	1	1	33	3	0	Firefox	Windows	Auckland	No
99	99	109	12752	311	50	Explorer	Windows	Auckland	Yes
2	2	2	779	16	0	Safari	Macintosh	Auckland	No
1	1	1	0	1	1	Firefox	Windows	Tauranga	No

Table 8 sample data set used for classifying

The decisions are typically straightforward attribute tests, employing one feature at a time to distinguish the data. New data can be categorized by sets of criteria defined at the nodes down. J.R. Quinlan (1993) has popularised the decision tree approach. The latest public domain implementation of Quinlan’s model is C4.5 [72]. We used 10-fold cross-validation. The data was divided arbitrarily into 10 parts in which the class was characterized in approximately the same proportions as in the full dataset. Each part was held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate was calculated on the holdout set. The learning procedure has performed a total of 10 times on different training sets, and finally, the 10 error rates were averaged to yield an overall error estimate.

We classified the users based on the browses and the city and we have found the users from Brisbane using Firefox didn’t buy but the users from the same city using Internet Explorer were most likely to buy.

All the users from Wollongong using Windows Chrome didn’t buy. The users from Whangarei using Windows

Chrome also didn't buy. The users from Auckland using Windows Chrome didn't buy but the users from the same city using Macintosh Chrome Internet Explorer were most likely to buy. From the same city, those using Firefox didn't buy but those who used Internet Explorer were most likely to buy. Also, those who used Safari Macintosh didn't buy. The users from Tauranga using Firefox didn't buy.

We can predict that all Internet Explorer users are more likely to buy and Google Chrome users will not buy. Therefore, it seems the websites don't work well with this browser or are slow to download, etc.

Table 9 shows the confusion matrix tree C4.5 decision tree classification using five features, 18,149 of the actual buyers' test set were detected as buyers. For precision, 0.97% was detected correctly.

Non-buyer	Buyer	Classified as
95587	1259	Non-buyer
1166	18149	Buyer

Table 9 Confusion matrix using decision tree classification by using 6 features

The detailed analysis of accuracy by J48 classification using six features is shown in Table 10. This has the advantage of generally improving system performance by removing inappropriate and unnecessary attributes.

TP Rate	FP Rate	Accuracy	Class
0.987	0.06	0.976	No
0.94	0.013	0.976	Yes
0.979	0.052	0.976	Avg

Table 10 Details accuracy by J48 classification using six features

1) Importance Attributes

The task of feature selection in this stage is to increase a performance condition such as accuracy and reduce the cost associated with producing the features. The reason for this is that most of the features may be redundant and inconsistent and affect the efficiency when data mining techniques are applied. A filtering approach will be used for this stage of feature selection because of the huge computational costs for the datasets. More precisely, information gain will be used as a feature selection to determine the ranking of input attributes and ranking the importance score for each attribute. The overall entropy (I) of a given dataset (S) is defined as [73]:

$$I(S) = -\sum_{i=1}^{c} p_i \, log_2 \, p_i$$

Where C means the entire amount of groups and pi the fraction of cases that fit class i. The decrease in entropy or the information gain is calculated for the individual feature in line with $\mathrm{IG}(\mathrm{S},\mathrm{A}) = \mathrm{I}(\mathrm{S}) - \sum_{\mathrm{v}\mathcal{E}\mathrm{A}} \frac{|\mathrm{S}_{\mathrm{A,v}}|}{|\mathrm{S}|} \mathrm{I}\left(\mathrm{S}_{\mathrm{v}}\right)$ where v a value of is A and $\mathrm{S}_{\mathrm{A,v}}$ the set of instances where A value has v. Figure 8 shows the results of ranking website elements by using Information Gain

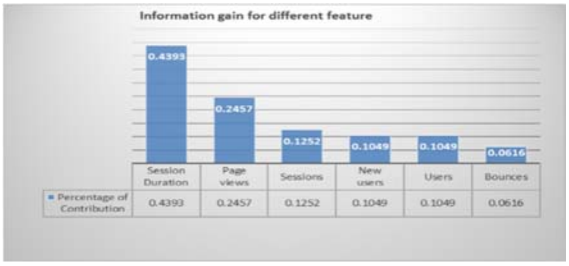

Figure 9 Ranking the website elements by using Information Gain

2) Identify visitor groups with common behaviour

Attributes	Cluster 1	Cluster 2
Users	1.5532	9.4361
Sessions	2.1869	14.5127
Session Duration	425.3737	5068.7619
Page views	8.9064	83.1075
Bounces	0.918	4.3056
Transaction	0	1.9139
Class	Non-buyer	Buyer

Table 11 Cluster Results

Table 11 shows the cluster results by applying the K-means algorithm. Each cluster shows us a type of behaviour in commercial website customers and from the cluster results.

F. Discussion

From the cluster results, we can begin to draw some conclusions. We have found that cluster 1, which means the users who did not buy, have a low number of sessions. This means the user looked at one page only. Therefore, the users are not interested to buy from this website so they leave straightaway. In the cluster 2 scenario, this means the users who actually buy, have a high number of sessions. This means the users are interested in buying and. they have more sessions on the websites. In other words, the web users didn't look at another page on the website within the next 30 minutes (that's how long a default session lasts).

For the attributes session duration, non-buyers have less session duration than the buyers, which is making sense as the buyer engaged more within the website or trying to find what they are looking to buy, while the non-buyers are not engaged with the website or they struggled to understand the website navigation scheme or to find meaningful content. For the page view attribute, the buyers have higher numbers of page views than the non-buyers; this is indicating the level of interaction and harmony between the users and website content. For the bounce attribute, although the website owners are trying to find methods to avoid or reduce the high bounces since they consider a high bounce rate is not a desirable outcome, in our case the buyers have high bounces with high conversion rates, which is a good indicator that the goal of the website has been achieved. In other words, the high bounce is not always a bad sign. Therefore, there is a strong relationship between bounce and purchase conversation rate. On the other hand, the non-buyers have a low bounce rate which may indicate that people who come to the website are not interested in buying. The bounce rate of a website can be explained in correspondence with the buying conversion rate, perhaps the customer has not yet decided to pay and is comparing the price in this website with other websites before purchasing.

The cluster attributes in this experiment are the same as the website issues and we can understand the session duration more importantly in terms of conversion than the bounces, so this also gives some guidance when I rank the issues.

IV. CONCLUSION

The Internet has become the world's main knowledge source. Extracting data from the web professionally is becoming gradually significant for various reasons. In this paper, the framework is proposed to understand user's behaviour on e-business websites. Data is extracted from several websites and converted to an understandable format. Then various data mining techniques are applied to this data. Our main finding is that the high bounces it is not always a bad indication as we found that buyers have higher bounce rates than the non-buyers. In addition, the users who have longer session duration, more page views and high bounces are a more likely purchase, while the users who have less number of sessions, shorter session duration, less number of page view and low bounces are less likely purchase. In addition, we have found the most important feature influencing the costumers' decision to purchase is the session duration and the least important feature is the bounces. Therefore, the webmaster team must concentrate on finding ways to keep users more engaged in the website in order to gain more conversion visitors.

References

[1] Statista. (2016, August). *Global number of digital buyers 2014-2019*. Available: http://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/

[2] E. Turban, D. King, J. K. Lee, T.-P. Liang, and D. C. Turban, *Electronic commerce: A managerial and social networks perspective*. Springer, 2015.

[3] F. B. Tan, L.-L. Tung, and Y. Xu, "A STUDY OF WEB-DESIGNERS'CRITERIA FOR EFFECTIVE BUSINESS-TO-CONSUMER (B2C) WEBSITES USING THE REPERTORY GRID TECHNIQUE," *Journal of Electronic Commerce Research*, vol. 10, no. 3, p. 155, 2009.

[4] W.-C. Chiou, C.-C. Lin, and C. Perng, "A strategic framework for website evaluation based on a review of the literature from 1995–2006," *Information & management*, vol. 47, no. 5, pp. 282-290, 2010.

[5] N. H. Borden, "The concept of the marketing mix," *Journal of advertising research*, vol. 4, no. 2, pp. 2-7, 1964.

[6] E. Constantinides, "The 4S web-marketing mix model," *Electronic commerce research and applications*, vol. 1, no. 1, pp. 57-76, 2002.

[7] J. W. Palmer, "Web site usability, design, and performance metrics," *Information systems research*, vol. 13, no. 2, pp. 151-167, 2002.

[8] P. Soonsawad, "Developing a new model for conversion rate optimization: A case study," *International Journal of Business and Management*, vol. 8, no. 10, p. 41, 2013.

[9] W. H. DeLone and E. R. McLean, "Information systems success: The quest for the dependent variable," *Information systems research*, vol. 3, no. 1, pp. 60-95, 1992.

[10] L. C. Schaupp, F. Bélanger, and W. Fan, "Examining the success of websites beyond e-commerce: An extension of the IS success model," *Journal of Computer Information Systems*, vol. 49, no. 4, pp. 42-52, 2009.

[11] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319-340, 1989.

[12] B. Clifton, *Advanced web metrics with Google Analytics*. John Wiley & Sons, 2012.

[13] Google. (2016, January, 2016). *Google Analytics*. Available: https://www.google.co.nz/analytics/

[14] L. Hasan, A. Morris, and S. Probets, "Using Google Analytics to evaluate the usability of e-commerce sites," in *Human centered design*: Springer, 2009, pp. 697-706.

[15] J. Weber, "Google Tag Manager and Google Analytics APIs," in *Practical Google Analytics and Google Tag Manager for Developers*: Springer, 2015, pp. 257-263.

[16] B. Plaza, "Google Analytics for measuring website performance," *Tourism Management*, vol. 32, no. 3, pp. 477-481, 2011.

[17] L. N. Carroll, A. P. Au, L. T. Detwiler, T.-c. Fu, I. S. Painter, and N. F. Abernethy, "Visualization and analytics tools for infectious disease epidemiology: A systematic review," *Journal of biomedical informatics*, vol. 51, pp. 287-298, 2014.

[18] H. Chen, R. H. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS quarterly*, vol. 36, no. 4, pp. 1165-1188, 2012.

[19] R. Kalakota and A. Whinston, "Electronic Commerce: A Manager's Guide Addison-Wesley, 1997," *Source: http://www. pitt. edu/~ gallet*

ta/commerce. html# materials[*http://wwwpersonal*](http://wwwpersonal)*. umich. edu/~ widmeyer/cis518/cis518-syllabus. html [Accessed July 20, 1998]*, 1996.

[20] C. A. M. Soares, *Applications of Data Mining in E-business and Finance*. Ios Press, 2008.

[21] H. Wang, C. Yang, and H. Zeng, "Design and Implementation of a Web Usage Mining Model Based On Upgrowth and Preflxspan," *Communications of the IIMA*, vol. 6, no. 2, p. 10, 2015.

[22] C. J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. García, "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11243-11249, 9/15/ 2012.

[23] J. Paul and J. Rana, "Consumer behavior and purchase intention for organic food," *Journal of Consumer Marketing*, vol. 29, no. 6, pp. 412-422, 2012.

[24] H. Xiaofeng, C. H. Q. Ding, Z. Hongyuan, and H. D. Simon, "Automatic topic identification using webpage clustering," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 2001, pp. 195-202.

[25] F. Ricci, L. Rokach, and B. Shapira, *Introduction to recommender systems handbook*. Springer, 2011.

[26] J.-K. Shang, W.-T. Hung, C.-F. Lo, and F.-C. Wang, "Ecommerce and hotel performance: three-stage DEA analysis," *The Service Industries Journal*, vol. 28, no. 4, pp. 529-540, 2008/05/01 2008.

[27] D. Thorleuchter and D. Van den Poel, "Predicting e-commerce company success by mining the text of its publicly-accessible website," *Expert Systems with Applications*, vol. 39, no. 17, pp. 13026-13034, 12/1/ 2012.

[28] Y. Yang, B. Pan, and H. Song, "Predicting Hotel Demand Using Destination Marketing Organization’s Web Traffic Data," *Journal of Travel Research*, vol. 53, no. 4, pp. 433-447, July 1, 2014 2014.

[29] M. P. Yadav, M. Feeroz, and V. K. Yadav, "Mining the customer behavior using web usage mining in e-commerce," in *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on*, 2012, pp. 1-5.

[30] E. Bigné ‐ Alcañiz, C. Ruiz ‐ Mafé, J. Aldás ‐ Manzano, and S. Sanz ‐ Blas, "Influence of online shopping information dependency and innovativeness on internet shopping adoption," *Online Information Review*, vol. 32, no. 5, pp. 648-667, 2008.

[31] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[32] J. K. Kim, Y. H. Cho, W. J. Kim, J. R. Kim, and J. H. Suh, "A personalized recommendation procedure for Internet shopping support," *Electronic commerce research and applications*, vol. 1, no. 3, pp. 301-313, 2003.

[33] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168-177: ACM.

[34] Y. Zhang and J. R. Jiao, "An associative classification-based recommendation system for personalization in B2C e-commerce applications," *Expert Systems with Applications*, vol. 33, no. 2, pp. 357-367, 2007.

[35] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web*: Springer, 2007, pp. 325-341.

[36] J. Yuantao and Y. Siqin, "Mining E-Commerce Data to Analyze the Target Customer Behavior," in *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, 2008, pp. 406-409.

[37] C. Yu and X. Ying, "Application of Data Mining Technology in E-Commerce," in *Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on*, 2009, vol. 1, pp. 291-293: IEEE.

[38] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Extracting Consumers Needs for New Products - A Web Mining Approach," in *Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on*, 2010, pp. 440-443.

[39] Z. Liu and L. Wang, "Study of data mining technology used for e-commerce," in *2010 Third International Conference on Intelligent Networks and Intelligent Systems*, 2010, pp. 509-512: IEEE.

[40] M. Zhang, "Research of personalization services in e-commerce site based on web data mining," in *2011 International Conference on Computational and Information Sciences*, 2011, pp. 438-441: IEEE.

[41] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing," *Expert Systems with Applications,* vol. 39, no. 3, pp. 2597-2605, 2/15/ 2012.

[42] W. Zuo and Q. Hua, "The application of web data mining in the electronic commerce," in *Intelligent Computation Technology and Automation (ICICTA), 2012 Fifth International Conference on*, 2012, pp. 337-339: IEEE.

[43] A. Rajaraman, J. D. Ullman, J. D. Ullman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press Cambridge, 2012.

[44] N. Poggi, V. Muthusamy, D. Carrera, and R. Khalaf, "Business process mining from e-commerce web logs," in *Business Process Management*: Springer, 2013, pp. 65-80.

[45] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Systems,* vol. 70, pp. 301-323, 2014.

[46] A. Alzua, J. K. Gerrikagoitia, and F. Rebón, "Tourism Destination Web Monitor: Beyond Web Analytics," *E-Review of Tourism Research (eRTR),* vol. 5, no. 5, 2014.

[47] N. Verma, D. Malhotra, M. Malhotra, and J. Singh, "E-commerce Website Ranking Using Semantic Web Mining and Neural Computing," *Procedia Computer Science,* vol. 45, pp. 42-51, 2015.

[48] U. Fayyad and P. Stolorz, "Data mining and KDD: Promise and challenges," *Future generation computer systems,* vol. 13, no. 2, pp. 99-115, 1997.

[49] K. G. M. M. Alberti and P. Z. f. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation," *Diabetic medicine,* vol. 15, no. 7, pp. 539-553, 1998.

[50] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," 1990.

[51] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, no. 1, pp. 81-106, 1986.

[52] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[53] L. Breiman, "Bagging predictors," *Machine Learning,* journal article vol. 24, no. 2, pp. 123-140, 1996.

[54] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[55] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 28, no. 1, pp. 100-108, 1979.

[56] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery,* vol. 2, no. 3, pp. 283-304, 1998.

[57] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[58] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electronic Markets,* vol. 26, no. 2, pp. 173-194, 2016.

[59] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management,* vol. 35, no. 2, pp. 137-144, 2015.

[60] R. Kosala and H. Blockeel, "Web mining research: A survey," *ACM Sigkdd Explorations Newsletter,* vol. 2, no. 1, pp. 1-15, 2000.

[61] O. Etzioni, "The World-Wide Web: quagmire or gold mine?," *Communications of the ACM,* vol. 39, no. 11, pp. 65-68, 1996.

[62] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[63] P. Lopes and B. Roy, "Dynamic Recommendation System Using Web Usage Mining for E-commerce Users," *Procedia Computer Science,* vol. 45, pp. 60-69, 2015.

[64] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *Acm Sigkdd Explorations Newsletter,* vol. 1, no. 2, pp. 12-23, 2000.

[65] B. Mobasher, "Web usage mining and personalization," *Practical Handbook of Internet Computing,* pp. 264-265, 2004.

[66] M. Perkowitz and O. Etzioni, "Towards adaptive Web sites: Conceptual framework and case study," *Artificial Intelligence,* vol. 118, no. 1–2, pp. 245-275, 4// 2000.

[67] S. Schechter, M. Krishnan, and M. D. Smith, "Using path profiles to predict HTTP requests,"

Computer Networks and ISDN Systems, vol. 30, no. 1–7, pp. 457-467, 4// 1998.

[68] G. Castellano, A. M. Fanelli, and M. A. Torsello, "Web Usage Mining: Discovering Usage Patterns for Web Applications," in *Advanced Techniques in Web Intelligence-2*: Springer, 2013, pp. 75-104.

[69] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM Transactions on Internet Technology (TOIT)*, vol. 3, no. 1, pp. 1-27, 2003.

[70] C. J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. García, "Web usage mining to improve the design of an e-commerce website: OrOliveSur. com," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11243-11249, 2012.

[71] Google. (2016). *Google Analytics*. Available: https://www.google.com/analytics/web/

[72] J. R. Quinlan, *C4. 5: programs for machine learning*. Morgan kaufmann, 1993.

[73] R. M. Gray, *Entropy and information theory*. Springer Verlag, 2010.