UNIVERSITY OF TWENTE.

Faculty of Behavioural, Management, and Social Science

Profiling of potential higher education website visitors based on online behaviours: A machine learning approach

Parth Gupta
Thesis Assignment
M.Sc. Business Administration – Strategic Marketing and Business Information
August 2018

Supervisors:
Dr. E. Constantinides (Efthymios)
Dr. S.A De Vries (Sjoerd)

Faculty of Behavioural, Management, and Social Science University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

Purpose: Recently, educational institutes have taken an initiative to aggregate and store the voluminous amount of behavioural data of users interactions on their websites, but still many have difficulties to unveil the patterns in it. Therefore, the objective of this paper is to discover the behavioural profiles of website users in the domain of higher education.

Design: In this research, a framework is developed for profiling of customer behavioural attributes within the marketing context. It define the process regarding the use of unsupervised machine learning algorithms in multiple stages for a variety of datasets which differs in terms of volume, ability to handle dimensionality, type (categorical/numeric) and its availability in R language. In addition, this study presented a model, which specifies the effect of the nature of information on the quality of clustering and difficulty to interpret them.

Findings: Outcomes from the application of unsupervised machine learning algorithms using the proposed framework on the Indian website visitors of University of Twente interested in master studies reveals that proposed combination and sequence of these algorithms performed well. These algorithms created the meaningful behavioural profiles as well as captured the minute differences between them.

Research limitations/implications: This study laid a foundation for future research work related to higher education website users in the domain of supervised machine learning especially classification. In this research, the behavioural profiles were discovered along with the patterns pertaining to each profile. Therefore, a prediction model can be built for Indian website visitors of the University of Twente interested in master studies to classify the new visitor belongs to this group in one of the six discovered behavioural profiles. Further, the text mining approach is suggested to unravel the semantics from the vast amount of unstructured text. This research is limited by volume and veracity of the dataset used.

Practical implications: The patterns in behavioural profiles render information to the marketer, to ameliorate targeting of the advertising campaigns by selecting a relevant target audience and messages based upon the behaviour manifested by visitors on the website. In the field of higher education unravelling patterns in the behavioural profiles will create possibilities for institutes to positively influence the engagement of their prospective students/users to aid them in their decision-making. It will simultaneously help the university to achieve their desired goals, such as improving the application submission rate. Further, it empowers the SMEs (Small and medium-sized enterprises) to efficiently execute the behavioural targeting under tight budget constraints or limited resources.

Originality/value: In this study, complete linkage (hierarchical clustering) followed by K-modes (partitional clustering) algorithms are executed using the proposed framework for behavioural profiling. To the best of the researcher's knowledge, in the domain of higher education, none of the studies used complete linkage (hierarchical clustering) in combination with K-modes to unravel patterns in behavioural attributes of website visitors. This methodology-oriented approach renders direction to create meaningful clusters for a small-scale symmetric binary dataset with low dimensionality.

Keywords: behavioural profiling, machine learning, nominal dataset, behavioural targeting

Paper Category: Research paper

Table of Content

List of Abbreviations

Abbreviation	Explanation
SET	Sustainable Energy Technology
ME	Mechanical Engineering
IDE	Industrial Design Engineering
IEM	Industrial Engineering and Management
BA	Business Administration
CE	Chemical Engineering
CEM	Civil Engineering and Management
ES	Embedded Systems
N	Nanotechnology
EE	Electrical Engineering
SE	Spatial Engineering
GSEO	Geo-information Science and Earth Observation
EEM	Environmental and Energy Management
HS	Health Sciences
MSM	Master Risk management
CS	Communication Studies
CME	Construction Management and Engineering
BIT	Business Information Technology
P	Psychology
EST	Educational Science and Technology
TM	Technical Medicine

1. Introduction

Empowered by the ever-rising employment of the internet and computing technologies, corporations gather massive amounts of consumer data, which is made feasible by advances in storage, networking, and data processing technologies. The swelling applications of neuroscience, internet of things, artificial intelligence, data mining, and social network analysis techniques have further fuelled the desire for personal information vis-à-vis effective strategic decision-making (Chester, 2012). These technologies are the primary input of customer data, which is used to personalize websites to surge the conversion rate (Tucker, 2014). Personal information is perceived as an ever more valuable commodity and the prevalent use of personal data for marketing purposes has created a market of it with an annual transaction of roughly 156 billion dollars (Montes et al., 2016). An industry group portrays online personalization as "the usage of technology and customer information to customise electronic commerce interactions between a business and each individual customer" (Adomavicius, 2006). Among marketers, personalization is usually presumed to be the utmost effective tool for attaining business success online (Cao and Li, 2007). Targeted advertising is a form of personalization in which advertisers target the individuals with customized-content; it has seen outstanding growth in the past few decades (Zhao, 2012; Zhao and Xue, 2013). Prevalent forms of this type of advertising are contextual targeting, behavioural targeting, IP-based geo-tracking and explicit profile data targeting (Lambrecht and Tucker 2013). A recent report states that digital advertising creates yearly revenue of €41.9 billion in Europe, soaring at a rate of 12.1 % year-on-year in 2016 (IHS Markit, "The economic value of behavioural targeting in digital advertising ", 2017). A rising percentage of this revenue and growth is specifically attributable to behavioural targeting. There are many definitions of behavioural targeting; it is also known as "behavioural profiling" and "online behavioral advertising" (Bennett, 2010). Instances include "a technology-driven advertising personalization method that enables advertisers to deliver highly relevant ad messages to individuals" (Nelson et al., 2016, p. 690) and "adjusting advertisements to previous online surfing behavior" (Van Noort et al., 2014, p. 15).

These definitions have two common characteristics; first, tracking or monitoring of user's online behaviour and second, utilize the gathered data to individually target ads. Therefore, Sophie et al. (2017) describe it as "the practice of monitoring people's online behavior and using the collected information to show people individually targeted advertisements". Online behaviour can include search histories, web-browsing data, media consumption, responses to advertisements, communication content and purchases (Zuiderveen Borgesius, 2015). The purpose of behavioural targeting varies across different type of firms, for instance, it was a world news that the American retailer Target knew about the pregnancy of a teenager before her father. Big-data (concerning the type of products and their frequency) was used by the Target to determine 'pregnancy prediction score' of women. If a woman scored higher than the benchmark set by the Target, they sent the discount coupons for the baby products.

Markets taking the lead in this domain are those with high advertisement expenditure per capita. Particularly in Europe, it includes Netherlands, France and UK, where behavioural data is utilized in more than 50% of entire digital display expenditure (IHS Markit, "The economic value of behavioural targeting in digital advertising ", 2017). In the eastern and southern European market, behavioural targeting varies between 5% to 20%, and it's even less if social media is excluded. This implies that behavioural targeting is still underutilized in most of the organisations. One of the obvious reason is that SMEs are overwhelmed by the variety, velocity and volume of consumer data in the contemporary situation. Secondly, paid services

to execute this task, is quite expensive. Eventually, due to a limited resource for advertising, SMEs end up with the inefficient usage of website visitor's data. Further, a recent scandal by Facebook (Social-media website) and Cambridge Analytics (British political consulting firm) to cunningly harness the personalized data of millions of users without their explicit consent to influence the aftermath of the Brexit vote and United States presidential election of 2016, significantly alleviates the concern of SMEs about handling of users' data (Meredith, 2018). Nevertheless, the dataset used in this study is anonymous and complies with GDPR regulations ("Guideline privacy rules: protection of personal data in scientific research", 2018).

Behavioural targeting has four essential stages (Srimani et al., 2011): first, uniquely recognize every user and generate a searchable database; second, track user's internet activities and record all relevant information (without personal identifiers); third, generate a user profile using an advanced algorithm and finally, send advertisements which are best suited to a specific individual based on past online behaviour. Although all the steps are crucial, the third stage- creating user profiles/ segments based on online behaviours- is the most critical and perplexing task. There is an enormous number of research articles available for disparate behavioural techniques and their contribution in ameliorating the online advertising regarding effectiveness and accuracy (Goldfarb & Tucker., 2011; Yan et al., 2009). However, there is a scarcity of research which outlines approaches for behavioural profiling based on the characteristics of data, unsupervised machine-learning algorithms and website visitor's behavioural attributes. Therefore, defining a framework for unsupervised machine learning algorithms based upon characteristics of online-behavioural data can empower small and medium-sized enterprises. Consequently, even with a low budget, SMEs can effectively employ behavioural targeting technique. A comprehensive understanding of these approaches will assist the firms to treat the data carefully and avoid the pitfall. Furthermore, the university considered in this paper has gathered the data about digital trails (behavioural attributes) left behind by visitors while surfing its website from past some years via Google Analytics and CRM software. Until now most of the patterns in the online behaviour of the users is unacknowledged to university (stated by the University marketing department). These unclear patterns present a challenge to dive deep into the huge piles of behavioural data of users to obtain patterns and analyse them from a disparate perspective to render meaningful insight to the marketing department of the university to design targeted campaigns or advertisements. Further, advertisements sent to the relevant cohort of users, will soar the user engagement and motivate/encourage the user to complete the conversion process. The conversion rate is the way of converting the website visitor into a paying customer. The usage of this term is conditional on the kind of websites, where some consider it as an outcome of actions other than sales ("Conversion Rate," 2017). An illustration of desired actions includes but is not limited to registration, newsletter subscriptions, submission of application dossiers.

The purpose of this study is to develop a framework for profiling of the behavioural attribute of customers, within the marketing context with the aid of machine learning algorithms. Furthermore, the objective is to unveil the patterns in the dataset to discover the behavioural profiles of website users of higher education. Since handling this mammoth data without restricting the scope of view can be a cumbersome and daunting task, therefore, this study will focus on exploring the patterns of behavioural data of website visitors of the University of Twente who are particularly keen in the Master of Science program. The results will unveil behavioural profiles of website visitors (interested in master studies) of the University of Twente and render insights about the online behavioural pattern of potential prospects.

Research Question: What are the behavioural profiles of Indian website visitors interested in the master studies at the University of Twente?

Further, following questions are necessary to be addressed to answer the main question

1. What are the characteristics of customer behavioural data used for profiling?
2. What sort of unsupervised machine-learning algorithm is appropriate for profiling nominal/categorical datasets?
3. How to identify the behavioural profiles?
4. Are the discovered behavioural profiles of Indian visitors consistent with an entire bunch of visitors interested in the master studies of the University of Twente?

To fulfil the objective of this study, relevant literature concerning segmentation such as machine learning and user profiling is reviewed. However, the core literature for this paper is user profiling, customer segmentation, clustering, unsupervised machine learning and identifying the suitable techniques for behavioural profiling of website visitors to augment the effectiveness of behavioural targeting.

The data used in this study are from the University of Twente, thus making it an explorative case study. Also, targeting relevant online advertisements to visitors will positively influence visitor's engagement, which may enhance, for instance, application submission rate, one of the crucial conversion points for most of the higher education institutes. Dataset used in this study is secondary, i.e. obtained from CRM (Customer Relationship Management) database and Google Analytics.

This study is an endeavour to close the gap in the literature, by developing the framework which consists of disparate machine learning approaches to leverage the customer attributes to support business decisions. Further, this study laid emphasis particularly on the symmetric binary dataset (category of the nominal dataset) of low volume and low dimensionality to execute unsupervised machine learning algorithms. This paper laid the foundation for future research in which analysis of the sequence of interactions on the university website along the timeline for the interested high potential prospect, i.e. 'Interested-HP' profile of Indian visitors interested in master studies could potentially reveal more insights to develop a robust prediction model, and further integration of text analytics could help to understand the semantics in behavioural data.

This paper is structured into five Chapters as follows. Chapter 2 is the theoretical framework which is the literature review of previous research on topics such as Knowledge Discovery in Datasets (KDD), Cross-Industry Standard Process for Data Mining (CRISP-DM), Machine learning, Behavioural targeting and user segmentation. Chapter 3 describes the methodology of this paper. It expands on the characteristic of the dataset and its collection method as well as analysis of a strategy, which machine-learning algorithm to execute. Chapter 4 outlines the outcome of all analysis, where analysis of each result is presented along with that results are visualised in a side-by-side fashion to support the interpretation. In Chapter 5, discussion and conclusion are illustrated in addition to limitations as well as theoretical and practical implication of this paper.

2. Theoretical Framework

To accomplish the objective of this paper, several aspects need to be accentuated. For instance, definition of behaviours, machine learning and its algorithm, customer characteristics etc. This chapter introduces the core literature used in this research. Every section illustrates the relevant facet of the key literature which permits the researcher to generate the framework consisting of ML algorithm strategies which is based on the attributes of data to execute clustering analysis with minimum inaccuracy. Relevant literature regarding each aspect is summarised and discussed briefly in each section.

2.1. Description of Behaviour

Prior to behavioural profiling, it is essential to understand and define the behaviour. Consequently, in this section, the definition of behaviour pertaining to this paper is expounded. In conventional terms, the behaviour is manner or fashion with which a being or system interact with one another. Cao (2014) states that behaviours are recognized by demeanour and actions with which beings interact with their environment. Behaviours had been immensely studied in the offline (non-digital) world from distinct viewpoints due to their explicitness (Cao, 2014). Though with the advancement of digital technology, behaviour took intricate forms, as it comprises the implicit form of digital information. For instance, the manner in which user seek out the information or respond to the digital environment. Behaviour logged in digital format are often termed "Soft Behaviour" or "Behaviour Computing" (Cao, 2014). In the field of behavioural informatics, Cao (2010) defines behaviour as "activities that present as actions, operations, events or sequences conducted by humans in specific context and environment in either virtual or physical organization". Behaviour computing is a favourable chance to ameliorate and discover certain behavioural patterns that could be utilized for distinct purposes in management and business intelligence (Cao, 2014).

Fayyad, Piatetsky-Shapiro & Smyth (1996) define the pattern as "an expression in some language describing a subset of the data or a model applicable to the subset". They accentuated that unravelled patterns is valid to some extent on new data that could render useful information which succour user in decision-making. Hence, Fayyad et al. (1996) concludes that for any pattern to be recognized as a pattern it has to exceed beyond a certain threshold to render meaningful information.

In brief, this paper considers the definition of behaviour defined by Cao (2010) as "activities that present as actions, operations, events or sequences conducted by humans in a specific context and the environment in either virtual or physical organization". In digital form, an illustration of behaviour in the context of this study comprises of actions (online-engagement) that users exhibit while surfing University website in order to harness the information. There are a bunch of behaviours (in either a virtual or physical environment) which demonstrates the behaviour of website users, which is used in this paper to represent the behavioural profile of users. To unearth the behavioural patterns of website users for higher education, specific techniques are needed to apply and extract significant insights. The following section summarizes the basics of such techniques (generally is known as *Knowledge Discovery* processes) which permits to extract information from the raw databases, which in this research is the behavioural data of website users.

2.2. Discovering Knowledge in Data

As stated in the earlier section, certain techniques are required which can be applied to the user's behavioural data to extract significant yet thoughtful informative insights. Data mining refers to the process of discerning meaningful patterns and trends in huge datasets. Big data and data mining go hand in hand. The challenges big data presents are frequently characterized by four V's– volume, velocity, veracity and variety. Volume is referred to the amount of data. Velocity represents the flow rate, i.e. the speed at which data is being generated and changed. Variety refers to distinct sorts of data being generated (clicks, numbers, text, etc.). Veracity refers to the issue of validity, meaning accuracy of data for the intended use.

Analysts from SPSS, Daimler- Chrysler, and NCR established the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Wirth et al.; 2000). CRISP-DM renders a non-proprietary and freely accessible standard process for fitting data mining into the generic problem-solving strategy of a research or business unit. As per CRISP-DM, a data mining project has a life cycle that consists of six phases - business/research understanding phase, data understanding phase, data preparation phase, modelling phase, evaluation phase and deployment phase, respectively. This phase-sequence is adaptive. That is, the next phase in the sequence often relies on the outcomes linked with the previous phase. The iterative nature of CRISP is represented by the outer circle in Figure 1.

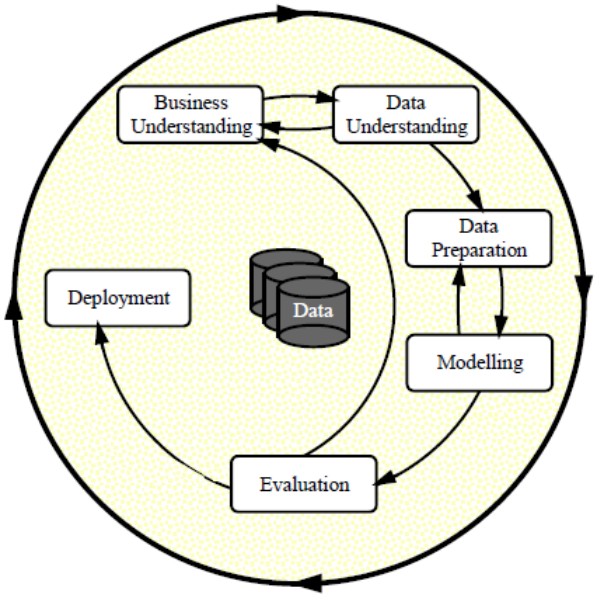


Figure 1: CRISP-DM Model for Data Mining. Reprinted from "CRISP-DM: Towards a standard process model for data mining," by R. Writh and J. Hipp, 2000, Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). Copyright 2000 by the DaimlerChrysler Research & Technology

Further, Fayyad et al. (1996) devised the term Knowledge discovery that consists of some previously mentioned techniques and distinguished it into two types: (1) Verification and (2) Discovery. The first category verification is limited to prove or disprove the hypothesis and second category Discovery, autonomously explore new patterns in data. Discovery is further subdivided into prediction and description. In prediction, the system finds patterns from behaviours (for instance number of clicks, content reading time) for predicting the future behaviour. On the contrary, descriptive sub-category unveils the naturally occurring patterns

in the dataset. In brief, descriptive category pertaining to the Discovery is relevant for this study. The descriptive technique can be used for clustering or segmentation, whereas, predictive technique can be utilized for customer characteristic prediction.

This section demonstrates the synopsis of data mining and its six phases of the life cycle and a description of techniques for discovering knowledge in data mining. Techniques which are usually used for a descriptive and predictive category of knowledge discovery are highlighted in the following section.

2.3. KD Modelling Techniques

This section describes the machine-learning algorithm used for subcategories of knowledge discovery, i.e. descriptive and predictive. Selecting descriptive and predictive one over another is entirely depends upon the research/ business goal of the individual. Machine learning field belongs to computer science that frequently uses statistical techniques which renders ability to computers to learn with the aid of data, without explicitly programmed. Nowadays, in the marketing domain, it is frequently used for customer profiling, classification and predictions. It is usually divided into two categories supervised-machine learning and unsupervised machine learning. The unsupervised machine learning resembles the descriptive category, as its algorithm identifies the naturally occurring patterns in the data whereas supervised-machine learning corresponds to the predictive category; it requires the output data, to train the input variables, to generate the classification/ prediction model. Supervised learning is occasionally called learning with a teacher; the teacher states the network which one is the accurate answer.

2.3.1. Unsupervised Machine Learning

Unsupervised machine learning is the task of deducing a function to describe the hidden structure from "unlabelled data". Clustering techniques are unsupervised machine learning, i.e. they do not require the target variable to identify arbitrary patterns in the data. It is an effective tool in managerial or scientific inquiry for discovering a natural pattern in the dataset. It splits a set of data into m-dimensional clusters, which are homogenous within and maximize the dissimilarity between them. Fahad et al. (2015) introduce a framework that categorizes the various clustering algorithms prevalent in the literature into distinct groups. These clustering algorithms can be broadly classified as follows: Partitioning-based, Hierarchical-based, Grid-based, Density-based and Model-based. Further, among these clustering algorithms in the marketing domain, three categories Partitioning-based, Hierarchical-based and Model-based are largely used in the scientific literature for segmentation (Fahad et al., 2015).

Hierarchical Clustering

In this clustering, data sets are arranged in a hierarchical manner, i.e. nested clusters, which can be organised as trees. The fundamental principle involves in this clustering is to examine the closeness of data points. Closeness is a subjective term, which can be selected from numerous definitions of similarity and distance measure (Pandove et al.; 2018). Hierarchical methods can either divisive or agglomerative. Agglomerative methods initiate with n clusters and successively merge similar cluster until the sole cluster is obtained. Divisive methods operate in the opposite direction, initiating with a sole cluster with includes all records.

Non-hierarchical/ Partitioning methods

A partitional or non-hierarchical clustering is a division of a set of observations into non-overlapping clusters (subsets) such that individual observation belongs to exactly one subset. Partitional clustering requires a predefined number of clusters to assign records to each cluster. These methods are generally computationally less intensive therefore it is preferred for large datasets. K-means is the prevalent clustering method in this category. There are numerous other partitioning algorithms such as PAM (Partitioning Around Medoids) and K-modes.

Model-Based Method

Such methods optimize the fit between the predefined mathematical model and given data. It is based on the supposition that the data is generated by a blend of underlying probability distribution. Further, it paves the way to automatically determine the number of clusters, which is based on standard statics and considering outlier, thus yields a robust clustering method. Self-Organising Maps (SOMs) or Kohonen maps is widely used model-based approach for clustering.

2.3.1.1. Literature Review of Unsupervised Machine Learning

Clustering is the key problem in data management and has an illustrious and rich history with literally thousands of disparate algorithms published in this domain. Even so, a sole method K-mean (partitioning clustering) remain the most prevalent clustering method; in fact; it was considered as one of the 10 algorithms in Data mining. Scaling, K-means to large datasets is relatively simple due to its iterative nature. However, the accuracy of the K-means procedure is highly dependent upon the choice of beginning seeds (Milligan & Cooper, 1987) and it often falls in local optima. Mishra et al. (2012) stated apart from local optima; the K-means algorithm is quite sensitive to the initial centroid. These are the major shortcomings of K-means, which decreases the accuracy of post-hoc (data-driven) market segmentation due to which the precise designation of market clusters is quite hard for marketing managers. To explore the solution to this problem the researcher reviewed the core literature related to the clustering and segmentation in the domain of marketing.

Punj and Steward (1983) suggested the combination of hierarchical approach i.e. Ward's minimum variance, followed by non-hierarchical approach i.e. K-means. His proposed algorithm provided better results than using either a hierarchical or a non-hierarchical method alone. Their approach is known as a two-stage approach. However hierarchical has its own disadvantages; first, for large datasets, it is expensive and slow; second, it cannot handle high dimensionality. Kuo et al. (2002) proposed a modified two-stage method, which initially uses the self-organising feature maps to determine the number of clusters and then employs the K-means algorithm to find the final solution. Pivotal steps taken by Punj and Steward (1983) and Kuo et al. (2002), revealed a good initiation point could significantly affect the K-means result and reduces the rate of misclassification. Therefore, Self-Organising Maps (SOMs) followed by K-means is a superior algorithm to obtain clusters.

Self-organising maps (Kohonen networks) well known for data visualization and clustering and it was formulated in 1982 by Teuvo Kohonen, which made him the most cited Finnish researcher. These neural networks have the ability to demonstrate the input signals as prototypes (weights), and with the visual examination of these maps, the resemblance between the observations can be inspected. The primary feature of the map is to preserve the original relationship between high-dimensional parameters while mapping them into low-dimensional maps; thus, a similar observation is likely placed in nearby regions (Kohonen et al., 2003).

These maps caught the attention of the researcher in different domains, ranging from biology to marketing (Augustijn et al., 2013). In its original version, the SOMs was aimed to cluster the real-valued data, so when it fed with categorical or binary data, the model usually delivers the worst results. Above that, when handling the categorical data other worries remains for example: encoding the categorical data into a real-valued vector and simultaneously conserving the categorical similarity information (Hsu, 2006). Some authors addressed the categorical/ binary learning problem via extending the SOMs to deal with these sorts of data. Lourenco et al. (2004) used the animal dataset to study different sort of similarity measurements, including euclidean distance, to examine the performance of Kohonen networks on binary data. But, still the unit's weight treats input value as real value. A similar incident happened in this research when the researcher tried to implement the SOM algorithm with euclidean distance on binary data. Appiah et al. (2012) introduced the tristate neuron weights strategy as a learning process, in which hamming distance was used instead of the euclidean distance. Trails with the MNIST dataset (LeCun et al., 1998) represented that; implementation of this approach is 30 times swift than original SOM, but it delivered the bad clustering accuracy. Santana et al. (2017) proposed effective SOM extension for the binary dataset, which considers both key training steps of self-organising map algorithm: determination/competition of winner neuron and network update rule. Santana et al. (2017) experiments with the proposed SOMs model deliver quite impressive clustering results as compare to other SOM models for binary dataset.

In this research, the researcher is using R language, which is free software for statistical computing and graphics. R and its libraries include clustering, classification and others machine learning algorithms. A key set of packages often included with the installation of R, and it has more than 12,500 additional packages as of May 2018 ("R (programming language)", 2018). However, there is no algorithm in R packages which can execute kohonen network algorithm on binary data. It is beyond the scope of research, to develop one for this research. In addition, K-means algorithm specifically performs best on metric dataset whereas produces meaningless clusters for categorical dataset because of integration of euclidean distance in its algorithm. Therefore, the researcher investigated for another set of techniques in the scientific domain to deal with binary/ categorical data.

2.3.1.2. Algorithm and similarity measures for Binary data

As mentioned in the previous section SOMs neither be used for dimensionality reduction nor to determine the number of clusters for binary dataset because of the inability of current packages of Kohonen network in R language to alter the original algorithm, which permits the use hamming distance instead of euclidean distance. There are other approaches available in the literature as well to explore the relationship between variables such as factor analysis or principal component analysis, (Jolliffe, 2002), or their categorical counterpart, correspondence analysis (Greenacre, 2010) is quite famous. However, the solution rendered by these methods is usually hard to interpret (Palla et al., 2012). There is a scarcity of approaches for categorical variable clustering, some of which are, for instance, hierarchical clustering or latent classes accentuated by Frolov et al. (2014) . Hierarchical clustering for the nominal, ordinal or metric dataset is often employed to determine the number of clusters. This sort of clustering based on proximity matrix, which includes dissimilates among all inspected variables. Dissimilarity can easily be computed by a simple transformation of the similarity measures.

Measuring distance or similarity between two data points is a key requirement for several data-mining tasks, which involves distance computation. Often, for continuous dataset Murkowski distance of order one, i.e. Manhattan distance and of order two, i.e. Euclidean

distance are two pervasively used distance measure for continuous data. The notion of distance for a categorical variable is not candid as for continuous data. The fundamental difference is that categorical attributes are not inherently ordered. If we consider two multivariate categorical data points, the distance or similarity among them is directly proportional to the number of characteristics in which they match.

(Boriah et al.; 2014) asked the question "Which similarity measure is best suited for my data mining task?", their experimental outcome suggested that there is no sole best performing similarity measure. Dataset used in this research is binary variables whose outcome can obtain only two values, i.e. 0 and 1. In nominal scale, observations are only allocated to different classes, but they can't be measured nor ordered. Binary variables often called symmetric, if there no specific choice for the outcome, i.e. both outcomes are equally valuable and assigned identical weight when proximity measure is calculated. On the contrary, if the outcome of binary variables is not equally valuable then, the binary variable is known as asymmetric. Tamasauskas et al. (2012) examined five asymmetric and five symmetric distance/ similarity measurements. Tamasauskas et al. (2012) categorised, Hamming, Dmatch, DSQmatch, Roger and Tanimoto and Sokal and Sneath 1 under symmetric distance measurement. Asymmetric distance measurement consists of Djaccard, Dice, Russell and Rao, Bray and Curtis and Kulcynski 1. Tamasauskas et al. (2012) chief aim was to evaluate the accuracy of distinct hierarchical clustering algorithms performance with respect to disparate similarity measures specifically for the binary dataset. Hierarchical clustering method used in Tamasauskas et al. (2012), research includes Average linkage, Centroid Linkage, Complete Linkage, Density Linkage, Flexible-beta, McQuitty's, Median, Single linkage, Two-stage density linkage and Ward's. Tamasauskas et al. (2012) experiment revealed that the symmetric distance measurements outperformed the asymmetric ones as their errors rate are smaller. Among the ten hierarchical algorithms in the study of Tamasauskas et al. (2012), complete linkage performed the best among all symmetric distance measurement whereas McQuitty's, Density linkage, Single linkage and Median algorithm performs the worst. Outcomes of the performance of hierarchical clustering algorithms on symmetric distance measurements are shown in Appendix 1.

Therefore, to determine the number of clusters (first stage of proposed framework section 2.9), researcher executes the hierarchical clustering with complete linkage algorithm and hamming distance (symmetric distance measurement).

Chan (2008) has categorized prevailing customer segmentation/ clustering methods into application-oriented and methodology-oriented approaches. Often methodology-driven studies modify data clustering techniques such as ward's minimum variance method, K-means, or use a blend of two or more data mining techniques to attain more accurate segments or clusters (such as Lee et al., 2004; Tzeng et al., 2007). "On the other hand, in application-oriented approach researcher must search for the optimum method for solving segmentation problems in specific applications" Chan (2008).

This study used a methodology-based approach. In methodology oriented, this research executes a combination of complete linkage (hierarchical clustering) followed by K-modes, to obtain the meaningful clusters for the symmetric binary dataset.

K-modes

Hung (1998) introduced K-modes, which is an extension of quite popular k-means procedure for continuous data to categorical data. Nevertheless, there are two key differences between these paradigms. First, since mean or average does not make sense for binary or categorical data, therefore the modal value of cluster is used; as the mean, the mode is also

considered component-wise. Second, instead of euclidean distance, K-mode uses hamming distance (Simple matching dissimilarity measure), again component-wise. Hung (1998) also stressed that; the K-mode clustering algorithm is swifter than K-means as it converges in less number of iteration. In principle, K-modes is like K-means, expect the two differences stated above. Like K-means, K-mode clustering algorithm requires the initial number of clusters, which difficult to determine and if determination of an optimum number of cluster went wrong, it would mislead the interpretation of results (Khan et al.; 2013). That is why the researcher adopted the two-stage clustering for the data set used in this research, i.e. firstly determine the number of the cluster via hierarchical clustering and then execute K-modes clustering algorithm.

The k-modes consists of the following steps (taken from Huang (1997)):

1. Select, the initial number of clusters k.
2. Allocate the data objects or observations to the cluster whose cluster centre closest to it.
3. Then, retest the dissimilarity of the observations against the current modes. If observation found nearer to the mode of another cluster then the current one, it reallocates observation of that cluster and update modes in both clusters.
4. Repeat step 3, until no observation has changed cluster membership.

In the upcoming sections, framework for multi-stage clustering is developed. It is designed to execute clustering on categorical and numeric data considering the volume and dimensionality of the dataset.

2.3.2. Supervised Machine Learning

Classification is a supervised machine-learning algorithm. The aim of this system is to generate a mapping (also called model) between a given set of documents and class labels. It is then used to determine the class of new unlabelled document automatically. To predict numerical variables (continuous values typically real value) a regression technique (linear regression or multiple regression) is used to approximate the outcome based on the new data. Another famous classification technique is decision trees, for variable which takes a discrete set of values.

This section accentuated numerous Knowledge Discovery techniques labelled by Fayyad et al. (1996), as Machine Learning methods in the IT field. Two main categories are supervised, and unsupervised machine learning, elaborate detail about each technique succours to recognize suitable technique to attain the purpose of this paper. However, further information is necessary how segmentation and user profiling should be done. Therefore, the following section summaries general approaches of segmentation in literature.

2.4. User profiling Approaches

To generate user profiles, understanding disparate approaches for segmentation, it is pivotal to strike an appropriate balance so that results are uncomplicated yet significant. Market segmentation includes a wide variety of approaches (Wedel & Kamakura, 2000). Fundamentally, these approaches can be categorised into two main cohorts. The first cohort ‘a priori’ is based on the known characteristics in advance (aware of the segmentation) to create useful grouping (common sense) before analysis is undertaken for instance socio-demographic characteristics (Boratto, Carta, Fenu, & Saia, 2016). The selection of personal attributes can be driven by practical consideration or experience with the local market. On the contrary, the

second cohort based on ‘post hoc’ or ‘data-driven’ method, i.e. empirical analysis is executed using multivariate analysis to recognize the segments. This approach creates a user profile, which is hard to interpret. However, it has the capability to divulge concealed relations among users that are unnoticed by the typical segmentation approach (a priori).

In an endeavour to alleviate the shortcoming of both approaches, Dolnicar (2004) suggests a *Hybrid* in which segmentation is done in two stages. The four approaches he introduced is a permutation of a priori and a posteriori method, i.e. a priori-a posteriori, a priori – a priori, a posteriori- a priori and a posteriori- a posteriori. For instance, in ‘a priori- a posteriori’ the segmentation process initiates with common-sense (a priori) segmentation and then every segment is separated into more refined sub-segment by using a posteriori (Dolničar, 2004).

In brief, there are three approaches to segmentation: ‘a priori’, ‘a posteriori’ and ‘hybrid’ (two-step approach). Among these approaches, Dolnicar (2004) proposed that hybrid approach balances the demerits of prior two approaches to achieve a robust segmentation, which is easy to interpret yet generates thoughtful insight. However, these approaches do not recognize how the nature of gathering user’s information effects the quality and interpretability of behavioural profiles. Therefore, the following section describes nature of gathering user’s information and user profiling methods.

2.5. User profiling Methods and Nature of Information

In prior section, approaches discussed for user segmentation were failed to consider the effect of nature of gathering user’s information on segmentation. In the literature, there are two central ways of obtaining information about the user. These are termed as an implicit or explicit information gathering. In the explicit method, the user explicitly provides information pertaining to the user’s preferences. The demerit of this method is that explicit profiles are static in nature and valid until the user alters their preferences and interest parameters explicitly. It is used for static profiling which analyses the static and predictable attributes of users. On the contrary, implicit information is harnessed dynamically by observing the user’s interactions with the system automatically. The implicitly generated profile is known as a dynamic or implicit user profile. Unlike static profiling, implicit profiling analyzes user’s behavioural pattern (e.g. past browsing behaviours) to identify user’s interests.

In implicit profiling, the precision of the user profile contingent on the volume of data generated through user-system interaction. It is also possible to create a hybrid user profile in two ways. One way is to initiate by explicit technique to gather data followed by an implicit technique. The second way is vice-versa. It has been cited that hybrid methods are more proficient than both aforementioned methods (Khosrowpour, 2005). Comparison of the aforementioned user profile type presented in Appendix 2.

In the study, researcher initially obtains explicit information about the prospect (website visitors) from CRM database (e.g. program in which website visitors are interested) then, implicit information from Google Analytics (e.g. type of device preferred by users). However, the proportion of implicit and explicit information can vary. Realizing nature of information for user profiling is not sufficient, understanding user-profiling methods render context to it and succours in their interpretation. There are essentially two types of user profiling method, which are collaborative and content-based methods. Khosrowpour (2005) proposed the third type, i.e. hybrid method, which is a combination of collaborative and content-based method. The content-based method also known as content-based filtering which assumes that user manifest the same specific behaviour under the same circumstances. Therefore, in this method

user's present behaviour is predicted based on the user's past behaviour. The system selects the things in which content correlation with the user profile is high. The content dependence is the key disadvantage of the content-based filtering. Also, these methods give poor performance if the volume of data is inadequate. Collaborative method also known as collaborative filtering method, which considers a user who belongs same cohort (e.g. same sex, age or social class) behave similarly and hence, have similar profiles. In this method users with similar taste, are referred to as 'like-minded people' (Cufoglu, 2014). The hybrid method integrates the advantages of both method and simultaneously eliminates the shortcomings. Summary of the user profiling methods proposed by Cufoglu (2014) can be found in Appendix 3. In this study profiling method is based on the Collaborative method.

In brief, there are three ways of obtaining user's information for profiling, among which hybrid information for profiling address the limitation of both implicit and explicit user data, which is highly pivotal for generating accurate user profiles. Further, in user-profiling methods, hybrid method tackles the shortcoming of a collaborative and content-based method to obtain clusters reflecting true behaviour of users.

2.6. Types of User-profiling and its Characteristics

In this section, various categories of data attributes used for customer segmentation/user profiling in the literature are mentioned. It is significant to recognize diverse customer data attributes, as utilizing a combination may yield a meaningful user profiles (from here segmentation and profiling are used interchangeably). Different data attributes of customers are used for different business situations. The following most widely used segmentation types and its characteristics are described:

Value-Based

Value-based data feature is used to classify the customers according to their value (Hosseni et al., 2011; Chorianopoulos et al., 2009). Customer value has been defined by numerous researchers, which usually recognised as customer profitability or customer equity (Hosseni et al., 2011). Relevant customer attributes pertaining to these are usage situation, necessities, favourite channel, preferable promotions etc.

Behavioural

Behavioural data feature permits the pattern recognition based on the passive behaviour of internet browsing session such as usage rate, visit frequency, benefit sought, the frequency of transactions, revenue history, user status (Baranowska, 2014). Hence, such a characteristic of data used to unveil patterns to group users with the same patterns.

Loyalty or Engagement

Loyalty or Engagement characteristics of data help to identify the distinct grouping of customers according to different extent of loyalty or engagement to brand or customer (Stroud, 2006). Characteristics related to these can be the frequency of purchases, the frequency of complaints, engagement scores, engagement interests etc.

Socio-demographics and Life Stage

Socio-demographics and Life Stage data characteristics aid in grouping the customers according to their social or demographic attributes. This segmentation is often used because attributes it comprises can influence the customer needs, preferences, attitude and usage behaviour (Chorianopoulos et al., 2009). Customer characteristics that can be followed and used for socio-demographic profiling are age, I.P (indicates location), gender, ethnicity, Facebook Id, Twitter Id , education other personal details.

Needs or attitudinal-based

These data characteristics succours to explore customer's needs which can be fulfilled by the purchase of a service or product, service, views, preferences and attitudes (Chorianopoulos et al., 2009). In the offline environment, the data required for this sort of profiling is fundamentally collected from external sources for instance market surveys via which customer can express their preferences and opinion. In an online setting, these characteristics consist of visited channels, visited Sites, page Views etc.

2.7. Data Sources

To obtain meaningful behavioural profiles, data sources of customer attribute as important as an appropriate clustering algorithm. The data sources for customer big data analytics mainly classified into five types: transactional data, data about service/product use, web behaviour data, data from customer-created texts and data about social network activities. This research is concentrated towards web behaviour about the user. Therefore, necessary data pertaining to it is extracted from web analytics. Instrumentation of web analytics implies that using methods and technologies with a purpose of recording and storing pertinent data connected to the interaction between the users and the system. Two distinct types of techniques applied when capturing user's behavioural data from an online services are page tagging and log files. The advantages and disadvantages of these two methods (page tagging and log files) are illustrated by Singal, Kohli, and Sharma (2014, p. 25).

2.7.1. Page Tagging (Google Analytics)

One general way to capture the user behaviour data from a website is the page tagging method. Page tagging depends on a piece of JavaScript that is injected to the source code of the page. When the browser loads a webpage, the browser executes the JavaScript code and data is transferred to the server hosting the analytics application. Depending on the need of the web site administrator, the tracking can be targeted only to a part of the website. Usually, all the pages of a given website are equipped with a tracking code. Applying page tagging to all the pages opened by the user basically enables the discovering the user behaviour more comprehensively as all the possible pages are being recorded. The Google Analytics is a widespread page-tagging tool as per W3Techs technology report nearly 86.4% of websites as per their database incorporated this in their web pages ("Usage statistics and market share of Google Analytics for websites", 2018).

2.7.2. Log files (CRM)

Log files denote the web server log files, which records the user interactions, for instance, pages opened by the users, in the form of a log file entry. In academia and literature, this technique is known as a transactional log analysis (Arshad & Ameen, 2015). CRM software's are often employed to obtain this type of data. As per the survey, nearly 55% of higher education institutes is not utilizing CRM data for enrolment and marketing purposes (Blackboard, 2014).

In brief, the above two methods of data collection, render the disparate type of details about the users. Analysing data sources and their scope gives deep insight to select sole or combination of data sources to achieve the purpose of this study. Next section briefly elaborates on the behavioural attributes utilized by previous studies.

2.8. Behavioural Attributes

As accentuated in the prior section, having a comprehensive understanding of user-profiling types renders relevant insight about behavioural attributes to achieve the purpose of this paper. There is numerous literature available in which researchers used behavioural attributes for user profiling (Chorianopoulos et al., 2009; Boratto et al., 2016). Behaviours which is often used in behavioural targeting are mentioned below (Pandey et al., 2011; Baranowska, 2014):

Interactions clicked	Offers viewed	Web-shop visits
Interactions viewed	Product type purchased	Operating System Version
Visited channels	Product type viewed	Campaigns
Visited sites	Number of visits	Referring URL
Page views	Number of page views	Location
Preferred social media	Average visit time	Sequence of page visited
Clicked banners	Membership	In-text semantics

As aforementioned, this study is based on fundamental assumption that user who belongs same cohort behave similarly (collaborative method) and vice-versa (Cufoglu, 2014).

2.9. Framework for User Profiling

In this section, the framework is presented, Table 1 describes the virtues and shortcomings of unsupervised machine-learning clustering algorithms in terms of size, the ability of handle dimensionality, noise, type of dataset and its availability in R language. Fahad et al., 2014 stated that size of data set has a considerable effect on the quality of clustering, i.e. some clustering algorithm is more efficient as compare to the other when the dataset is small and vice-versa. Datasets with numerous attributes are denoted as high dimensional because handling them presents the specific computational challenges. Fahad et al., 2014 also mentioned, handling high dimensionality is an essential feature in clustering analysis as numerous application requires the analysis of object with a high number of attributes, for instance, a text document may contain hundreds of keywords as attributes. However, it is difficult due to the curse of high dimensionality as some dimension are not relevant in the dataset, and it increases sparsity thus make interpenetration of clusters meaningless. Regarding the curse of dimensionality, the pivotal problem lies in the loss of discriminative power of density or distance measure (Assent, 2012). Assent (2012) did a comprehensive review of the clustering of high dimensional datasets. However, Assent (2012) found that there is no persistent definition in the literature pertaining to the minimum of dimensions, which can be considered as a high dimension. In some studies, data with as few as ten variables (dimensions) are referred as high dimensional, whereas numerous works, specifically in image processing, bioinformatics have hundred, or thousands of attributes (see, e.g., Jiang et al., 2004; and Kailing et al., 2003). Therefore, for this study researcher is considering <=10 as low dimension and >10 high dimension.

In Chapter 2 section 2.3.1, the researcher discussed the importance of executing two-step clustering and accentuated how machine learning algorithm counters each other merits

and demerits in particular scenario to obtain simple yet robust cluster which will render significant insights. Therefore, the framework in Table 2 represents the strategies under what condition which combination yields near to optimal results. The first stage is designed to determine the number of clusters. Therefore, all the algorithms (hierarchical or model-based) mentioned in stage 1 for the numerical or categorical dataset, would explore certain patterns and identify the suitable number of clusters. Stage 2 algorithms require the number of clusters to execute the final clustering. K-means, K-modes, SOMs/ Kohonen maps and traditional hierarchical clustering methods such as ward's minimum variance, complete linkage, single linkage etc. already discussed in Chapter 2 section 2.3.1. This section renders very short introduces to ROCK, CURE and CHAMELEON algorithm (Pandove, 2018).

ROCK (Robust Clustering using links): This clustering algorithm belongs to the agglomerative hierarchical clustering algorithms. It works well on both boolean and categorical variables, and it uses the concept of links to measure the likeness between a pair of data points (Guha et al.; 2000).

CURE (Clustering Using Well Scattered Representatives): It is used to illustrate the clusters with the aid of well-distributed representative points. Cluster distance in these techniques is the minimum distance between the representative points, which implies this incorporates both average and single linkage methodologies (Guha et al.; 1998). It is cables of capturing cluster of arbitrary shapes by selecting the scatter plot.

CHAMELEON: It measures the similarity between clusters based on the dynamic model. Fundamentally, in clustering process, two clusters can be collated only if closeness (proximity) and inter-connectivity between the clusters are similar to the internal-connectivity of the clusters as well as the closeness of items within the clusters. The methodology of dynamic modelling of clusters utilize in the CHAMELEON is valid for all type of dataset providing similarity matrix can be constructed (Karypis, 1999).

Table 1: Illustrates virtues and shortcomings of unsupervised machine learning algorithms

Clustering ML Types	Clustering ML Name	Size of Data	Handling High Dimensionality	Handling Noisy Data	Type of Data Set	Availability in R
Partitional Algorithm	K-means (MacQueen, 1967)	Large	No	No	Numeical	Yes
	K-Modes (Huang, 1997)	Large	Yes	No	Categorical	Yes
	K-Medoids (Park et al; 2009)	Small	Yes	Yes	Categorical	Yes
Hierarchial Algorithm	CURE-Clustering Using Representative (Guha et al; 1998)	Large	Yes	Yes	Numerical	No
	ROCK-Robust Clustering Algorithm (Guha et al; 2000)	Large	No	No	Numeical/ Categorical	Yes
	Chameleon (Karypis, 1999)	Large	Yes	No	Numeical/ Categorical	Yes
Model-based Algorithm	SOMs (Kohonen, 1998)	Small	Yes	No	Mutivariate Data	Yes

Table 2: Framework outlining various strategies for user profiling based on characteristics of dataset and unsupervised machine learning

Numerical Dataset		Unsupervised ML Algorithms	
Size	**Dimensions**	**Stage 1**	**Stage 2**
Small (<=2400 observations)	Low (<=10)	SOMs	K-Mean
	High (>10)	SOMs	K-Mean
Large (>2400 observations)	Low (< 10)	CURE	K-Mean
	High (>=10)	CURE	K-Mean
Categorical Dataset		**Unsupervised ML Algorithms**	
Size	**Dimensions**	**Stage 1**	**Stage 2**
Small (<=2400 observations)	Low (< =10)	Ward's Method\Complete Linkage	K-Mediods/K-Modes
	High (>10)	Chameleon	K-Mediods/K-Modes
Large (>2400 observations)	Low (< =10)	ROCK	K-Modes
	High (>10)	Chameleon\ ROCK	K-Modes

2.10. Model for determining the quality and interpretability of User Profiling

In section 2.4 and 2.5 of Chapter 2 approaches to user profiling and nature of data sets pertaining to the user's information was discussed. Cufoglu (2014) sated two ways of obtaining information about users, i.e. implicit and explicit information gathering. In the explicit method, the user of the system explicitly renders information related to his/her preferences. The demerit of this method that explicit profiles are static in nature and valid until only the user alters their preferences and interest parameters explicitly (Gena, 2005). It is used for static profiling which analyses the static and predictable attributes of users. On the other hand, implicit information is harnessed dynamically by observing the users interactions with the system automatically. The implicitly generated profile is known as a dynamic or implicit user profile. Unlike static profiling, implicit profiling analysis user's behavioural pattern (e.g. past browsing behaviour) to identify user's interests (Khosrowpour, 2005). Here, the precision of the user profile contingent on volume of data generated through user-system interaction. It is also possible to create a hybrid user profile in two ways. One way is to initiate by explicit technique to gather data followed by an implicit technique. Khosrowpour (2005) stated that the hybrid method is more efficient in comparison with the sole implicit or explicit method. Cufoglu (2014) accentuated the accuracy of implicit information ameliorate with continuous use of the system by users whereas the accuracy of the explicit/personalized information depends upon the manually rendered information which is updated by the user. Example of implicit (static) information consists of native country of the visitor. On the contrary, dynamic information consists of needs, behavioural attribute etc.

Table 3: Model to determine the quality and interpretability of user profiling based upon the nature of information

Nature of information	Quality of profiling (in terms of accuracy)	Efforts in gathering data	Difficulty in interpreting
Explicit Data (static information)	Contingent on the accuracy of information provided by the user	High	Low
Implicit Data (dynamic information)	Contingent on the amount of interaction between user and system (Higher interaction higher accuracy)	Minimal effort by the user	High
Implicit and Explicit Data	Contingent to both stated above	Moderate	Moderate

2.11. A literature review of other techniques implemented in the domain of Behavioural Targeting for User-profiling

This section briefly discusses the prior research, in the domain of customer profiling and illustrates the various techniques used to achieve this. Further, it will enhance the span of understanding and render indication which works has been tried and tested so far. Researcher discovers fundamentally; there are two categories of research in user profiling. In the first category, researchers proposed modified algorithms in terms of effectiveness, such techniques mentioned in the prior section**.** In the second category, the researcher introduced approaches

for user profiling which are either novel or first time implemented in the domain of User-profiling. Such approaches are mentioned below:

Low et al. (2011) addressed the problem related to the creation of a user profile for behavioural targeting. They demonstrated, time altering the hierarchical user model which harness the short and long-term user interests and this model is known for its application at web scale. They represented that streaming distributed inference algorithm can handle tens of millions user data and simultaneously accustomed to user's interest as it aids to know about the user. Their learnt user demonstration was empirically proven effective in computational advertising.

Yao et al. (2010) proposed a two-step approach for profiling the users, i.e. by integrating SOM-Ward clustering with predictive analytics. Yao et al. (2010) mentioned SOM-Ward Clustering is based on the unsupervised neural clusters, thus useful for exploratory data analysis, i.e. when no a priori categories have been identified, and it possesses robust capabilities in handling missing data, non-linear relationships and skewed distributions. On the other hand, predictive customized to a specific purpose and it is a part of supervised machine learning. His results demonstrate this combination is potentially effective in customer profiling/ segmentation.

Gong et al. (2013) argued that often user profiling work adapts clustering algorithms and treat data as user reprentation, but it does not consider the concealed semantics embedded in data. Therefore, in their research paper, they presented the behaviour based latent semantic user profiling and authenticate its validity on new ads. They illustrate data as characteristics of user issued query or clicked advertisements after that Latent Dirichlet Allocation (LDA) is applied to the dataset for user profiling. Gong et al. (2013) found as compared to famous k-means clustering their approach attain higher Click- through rate (CTR) values on advertisements.

Tu et al. (2010) also profiled the users with the aid of Latent Dirichlet Allocation (LDA) and compared with other three baseline user profiling approaches hierarchical clustering, k-means and pLSA (Probabilistic Latent Semantic Analysis). Their results demonstrate that the proposed method delivers better performance as compared to other baseline user profiling algorithms, both in terms of assessment metrics click entropy and CTR improvement.

Wu et al. (2009) established a novel approach known as pLSA (Probabilistic Latent Semantic Analysis) for user profiling and they compared with conventional user profiling methods. Their empirical results showed that pLSA is better as compared to other techniques in terms of CTR enhancement. Boratto et al. (2016) present another novel method in which he profiled the users by analysing the things positively evaluated by users to obtain reliable user preferences. Then, words are extracted via word embedding's and novel class model, which authors named as neural class embedding. This model aids in understanding and categorizes every class of item to create profiles with a certain characteristic, thus avoiding generation of trivial segments. Further, they argued, this method renders easy interpretability of profiles. In the next section, the methodology is described to accomplish the purpose of this study.

3. Methodology

This chapter expounds the procedure to systematically realize the goal of this paper. The procedure is based upon the data-mining framework CRISP-DM (Cross-industry standard process for data mining), which outlines 6 phases of the data mining life cycle, i.e. business/research understanding phase, data understanding phase, data preparation phase, modelling phase, evaluation phase and deployment phase, respectively. This phase-sequence is adaptive. That is, the next phase in the sequence often relies on the outcomes linked with the previous phase. This chapter explores the first four stages of CRISP-DM: research-understanding phase, data understanding phase, data preparation phase and modelling phase. Further, the strategies are incorporated from the proposed framework and a model in the CRISP-DM framework for reliable user segmentation.

3.1. Research Understanding Phase

In this phase, the objectives of the research are coherently stated. As per the figures on the university website ("FACTS & FIGURES", 2018), the total number of students took admission is higher education studies of University of Twente was 10026 in 2016 out of which 53.06% are bachelor students, 41.15% are masters students and rest 4.15% are pre-master and post-master students. Admission criteria for these categories of students are distinct from each other therefore the researcher is anticipating, they reflect different behaviours while surfing the university website. Further, there are 2500 international students from nearly 80 nationalities, i.e. they form approximately 25% of the admitted students, which is interesting figure because although they form one-fourth of the total admitted students but their tuition fee (€ 11,500 / € 14,250 per annum for master studies , Non-European international students) is 5-6 times more as compared to national students or EU international students (€ 2,060 per year for a master studies) which implies that solely in terms of monetary value Non-European international students are highly beneficial. To encourage them to submit their application, it is pivotal to know their online-behaviour. However, the due presence of large number of nationalities, it is nearly impossible to analyse all of them within a limited time. Therefore, this study is focused on countries from where the majority of the traffic of prospective students arrive at the website of the University of Twente. India and Germany are the two countries that have the majority of prospective student traffic after the Netherlands. Therefore, the study/research focuses on prospective Indian master's students/ website visitors. To make the study relevant a two-step analysis was executed: first, analyse patterns in the behavioural data of All Master visitors. In the second step, analyse behavioural patterns of Indian Master visitors and compare it with All Master visitors. The results would preferably lead to the discovery of behaviour manifested by potential prospects.

3.2. Data Understanding Phase

It elaborates about the nature and method of data collection. The data used for this study is secondary in nature, i.e. it is extracted from the existing database of the university. Further, it exhibits both implicit and explicit characteristics, for instance, data extracted from page tagging is an implicit form of user data (downloaded PDF from the website) and the type of study in which user is interested illustrates the explicit user data. Cufoglu (2014) mentioned that if the dataset is a combination of explicit and implicit information then it renders high quality user profiles as compared to sole implicit or explicit information (Chapter 2, section 2.5). Since data is highly sensitive, therefore, it is handled carefully to prevent breaching of the

visitor's privacy. Entire data is anonymised, and it is harnessed with the clear consent of the user. It meets all the requirements of General Data Protection Regulation ("Guideline privacy rules: protection of personal data in scientific research", 2018).

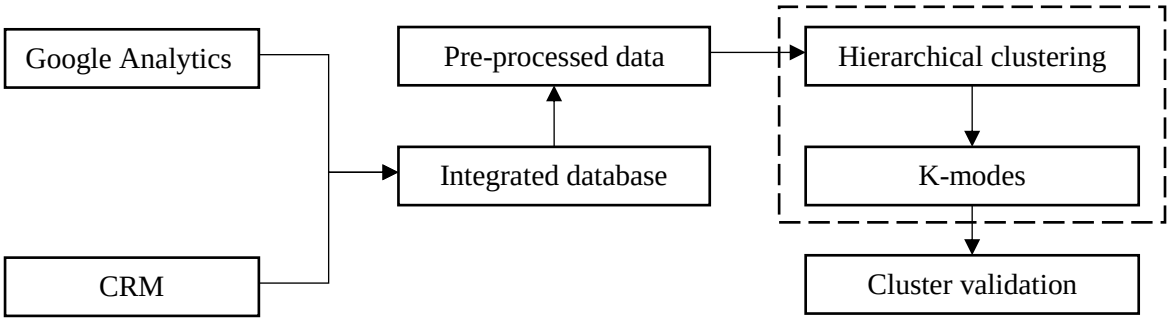
Two sources are used for obtaining behavioural data; Google Analytics (page tagging) and CRM database of the university. These datasets are merged to create the dataset which has both implicit and explicit characteristics, but mostly explicit. However, data is of a secondary nature, its reliability and quality are often questioned in research. Therefore, to ameliorate the quality and reliability, validation of the data is necessary.

3.3. Data Preparation Phase

This labour-intensive stage covers all aspects for formulating the final dataset, from the raw and dirty datasets. Further, it represents the approaches of analysis implemented in this paper. It renders the comprehensible understanding of required steps, which can be replicated for the new data. To achieve the purpose of this paper analysis are conducted using open-source software R and Microsoft Excel, which was purposefully chosen as one of the objectives of this study to enlighten the SMEs about the cheap alternatives for advanced clustering.

Figure 2 represents all the required steps to execute all analysis in this study. It initiates by retrieving data from the university account of Google Analytics such that it must have wrd-ID corresponding to extracted behaviours. wrd-ID is the anonymous user-ID which is generated when visitor first time visits the university website, and it is the only link which can integrate disparate database. Excel files with behavioural attributes are extracted multiple times from Google Analytics from Oct 2016 until July 2017, in a step of one month to get a high sampling rate (83.5%). After that, relevant behavioural features extracted from multiple CRM excel files of the University of Twente. Then, these two data sources integrated into one comprehensive file, followed by pre-processing of data set in suitable machine learning format by transforming the categorical variables into dummy variables.

Figure 2: Demonstration of steps for conducting the analysis



3.4. Modelling Phase

In this phase, the researcher must select and execute the appropriate clustering algorithm. The volume of pre-possessed data is low volume as well as dimensionality (2327 entries and 11 behavioural dimensions), therefore, as accentuated by the proposed two-stage clustering framework in Chapter 2 section 2.9, complete linkage (hierarchical) clustering with hamming distance followed by K-modes is the optimal choice. There are two analyses conducted in this paper as mentioned before. Procedure and selected strategies for each analysis are demonstrated below. The first analysis is executed on All Master visitors to build the

comprehensive view from the discovered profiles. The procedure for executing analysis is shown in Figure 3. The analysis is conducted by using implicit and explicit data with one-step a postpriori approach (refer to section 2.4 and 2.5). To achieve the desired purpose, i.e. behavioural profiling of Indian Master visitors, the researcher introduced the variable country. Therefore, one more analysis will be executed using the proposed framework and model in this paper.

Figure 3: Demonstration of steps taken for first analysis

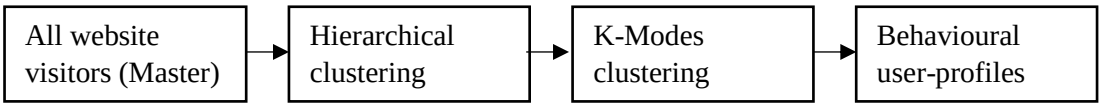
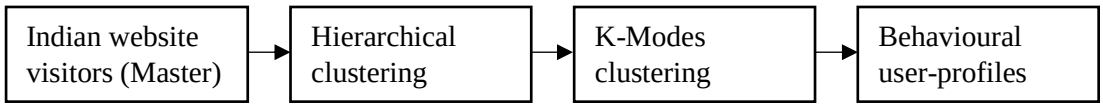


Figure 4, represents the second analysis, conducted on Indian Master visitors, with implicit and explicit datasets along with a prioi- a postpriori approach.

Figure 4: Demonstration of steps taken in second analysis



The profiling results of the above two analysis are elaborated in terms of behaviour exhibited by each profile, which are presented in the next chapter. To assess the reliability of behavioural profiles, the results are validated by silhouette analysis.

3.4.1 Behavioural Attributes

The purpose of this study is to find behavioural profiles of Indian visitors who were interested in the master studies. Marketing and Communication department of the University of Twente acknowledge that some student initiates their application early. Therefore, it is necessary to motivate them until they finish the application submission. Behavioural patterns of prospective student help in predicting other behaviours which user can exhibit at later stage. Therefore, it will help marketing department to put forth advertisements strategically, which will compensate for the deficiency in knowledge/skill of the user, thus enhances user engagement and motivates them to submit the application dossier. Hence, application submission is the explanatory variable for this study.

The pre-processing of data generated numerous variables, which can indicate, what sort of online-behaviour influence the Indian prospects to submit the application dossier. There are ten behavioural features, which are extracted from pre-processed data which are list below and rest of the variables (country, continent, device type used and source) helps in further segmentation (micro-segmentation) to refine the targeting of advertisements.

Osiris Application Submitted	Education brochure request	Question via web form
Managed CTA click	Registration Open day	Request student for a day
PDF download	FAQs	Eligibility check
Scholarship finder		

The first feature Osiris application submitted is manifested when prospective visitor submits their final application via university online platform Osiris; this is the conversion point (Explanatory variable) for this study. The Managed CTA (click) is when the user first interacts with the website and return to the University website, the call to action content and message is tailored to motivate or encourage the user to act. Next, behavioural feature PDF download is when a user downloads the non-study related brochure. Scholarship finder is the behaviour that user manifest when the user at least once went to scholarship webpage and searched for the scholarship. After that education brochures request is when a user at least once requested for education brochure. Open day registration behaviour represents that the user wants to experience the study and culture of the university. FAQs is the behaviour that user illustrates when user went to the FAQs webpage, to resolve general doubts and questions. Next, a question via web form when the user asked a question via web form, and this feature is available on the web pages of the university website. Next, Eligibility check is when the user went through the procedure to know whether he/she is eligible for the study program of their choice. Request student for a day is behaviour, which illustrates the user willingness to be a student for one day to experience the study and campus.

4. Results

In this Chapter, the proposed framework for clustering of the categorical dataset and a model is implemented to analyse the behavioural profiles of the website visitors of the University of Twente. The outcome of the analysis is illustrated in the form of tables or figures. Furthermore, the discovered behavioural profiles of website visitors of the University of Twente who are interested in the master studies are described.

The pre-processed dataset has the behavioural attributes of 2327 users, who are keen in the master's level of study. The range of behavioural attributes is one; this is because all categorical features are converted into binary features. Each binary feature has a minimum value of zero and the maximum value is one. The zero corresponds to the absence of a characteristic, and one corresponds to the presence of behavioural characteristic. There not much significance of minimum and maximum value in this descriptive statistics, besides it renders a warning if there is an error in data. Table 4 indicates that there is no irregularity in the dataset considering range, minimum and maximum values.

Figure 4: Descriptive Statistics of pre-processed database

Descriptive Statistics of pre-processed database

Behavioural Attributes	*N*	*Range*	*Minimum*	*Maximum*	*Mean*
Managed_CTA_Display	2327	1	0	1	.81
Managed_CTA_Click	2327	1	0	1	.28
PDF_download	2327	1	0	1	.16
Scholarship_finder	2327	1	0	1	.11
Education_broucher_request	2327	1	0	1	.59
Registration_Open_Day	2327	1	0	1	.08
FAQs	2327	1	0	1	.03
Question via Webform	2327	1	0	1	.04
Request student for the day	2327	1	0	1	.09
Eligibility Check	2327	1	0	1	.28
Osiris_Application_Submitted	2327	1	0	1	.12
Valid N (listwise)	2327				

Dataset also contains four factors along which these behavioural attributes can be described more precisely, i.e. country, device type, program and source to the website (the origin of the website traffic, e.g. google.com). The mean of each behavioural attribute in Table 4 represents the frequency percentage of 2327 visitors dataset for instance approximately 28% went through managed CTA click, nearly 59% requested for education brochures, around 28% went through eligibility check, and roughly 3% went to frequently asked questions webpage which is nearly tantamount to 70 users.

Dolincar (2008) mentions that a number of variables that can be evaluated with a sample of a certain volume are limited and there is no particular rule for non-parametric techniques. However, a rule of thumb suggested by Formann (1984) renders some guidance: for the situation of the binary dataset, the minimum amount of observation must be 2^k (K= number of variables), preferably $5*2^k$ of observations. Since a number of behavioural attributes are 10, therefore, the number of observation as per minimum criteria is 1024 and as per the preferred 5120 criteria. Consequently, the size of data is voluminous enough to fulfil the pre-

condition of clustering analysis (2327). Further, pre-processed data do not have any outliers; it does not require standardisation due to the binary characteristics of data which implies distance and range of all behavioural attributes are alike. Hence, the dataset is in the right condition to proceed with analysis. As discussed in Chapter 3, the analysis will be conducted in two stages; first, determining the number of clusters via hierarchical clustering followed by K-mode clustering.

4.2. Calculating the number of clusters

To decide an optimal number of clusters for a symmetric binary dataset hierarchical clustering method is being adopted. As discussed in Chapter 2, Tamasauskas et al. (2012) evaluated performance (in terms of accuracy) of distinct hierarchical clustering algorithms (ten) with respect to disparate symmetric and asymmetric similarity measures specifically for the binary dataset. The outcome of the study (Tamasauskas et al.; 2012) demonstrated that complete linkage hierarchical clustering methods with symmetric similarity metric performed the best as compared to all other hierarchical algorithms considered in the study.

Hence, to identify the initial number of clusters; the researcher executed the hierarchical clustering with complete linkage hierarchical algorithm and hamming distance (symmetric distance measurement). Further, Figure 6 and 8 represents the best partition to cut dendrogram based on high relative loss of inertia method. This method was primarily proposed by the HCPC function (Hierarchical Clustering on Principal Components) of the package FactoMineR (Husson et al.; 2018). In this method sum of within-cluster inertia is computed for each partition. The best partition is highlighted in black and the second-best is accentuated in grey.

Figure 5 represents the dendrogram of all master prospects and Figure 6 manifests an appropriate number of clusters for all master prospects. The four clusters recommended by the method is highlighted in black point and second best accentuated in grey point, which is three clusters. However, in dendrogram, we can visually inspect that six is an appropriate number for clustering. Further, the difference between the relative loss for 3 clusters and 6 clusters is nearly 0.006 which is negligible. Dendrogram for Indian prospects is demonstrated in Figure 7 and Figure 8 represents three clusters are the best choice for Indian prospects, and the second-best choice is six clusters. If we visually inspect the dendrogram of Indian prospects, 6 clusters for Indian prospects appears to be appropriate.

Figure 5: Dendogram of All Master visitors

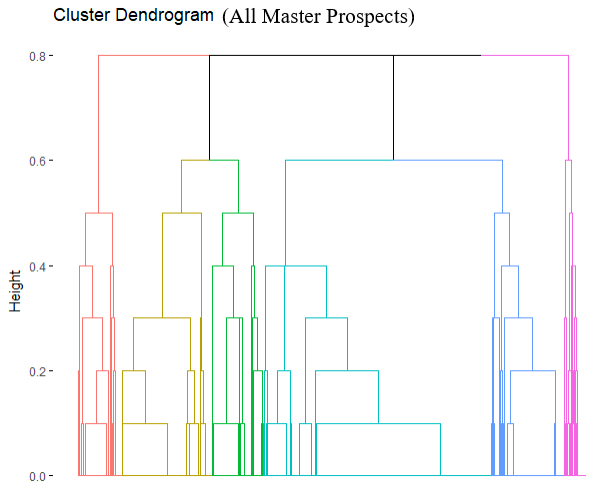


Figure 6: Optimal number of clusters for All Master visitors

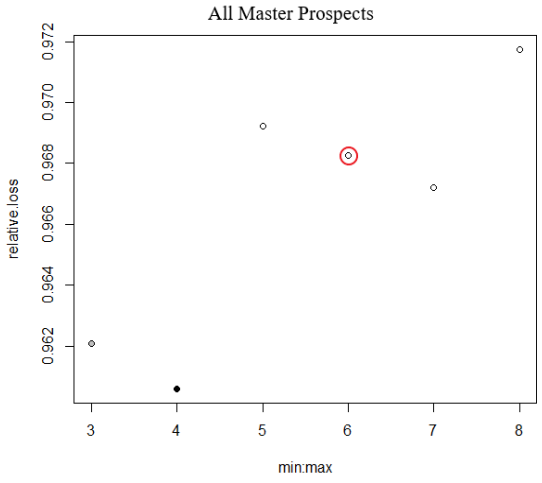


Figure 7: Dendogram of Indian Master visitors

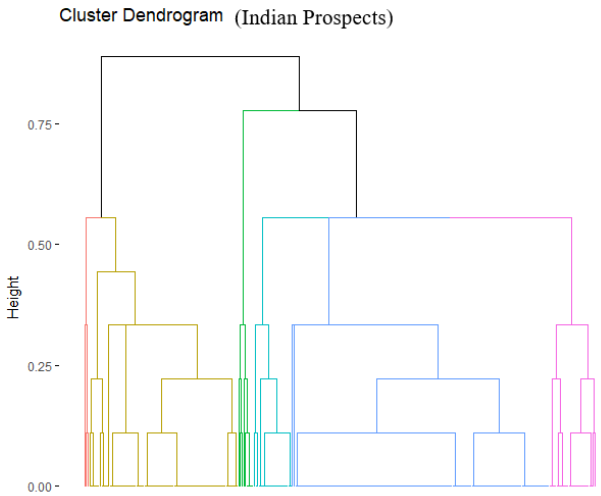
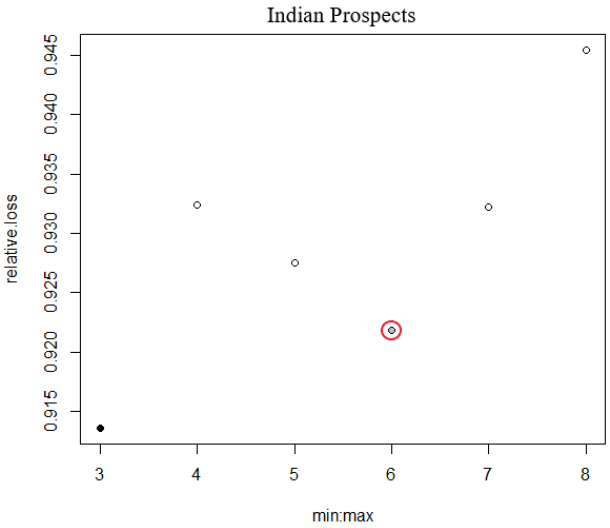


Figure 8: Optimal number of clusters for Indian Master visitors



Hence, the appropriate number of the profile using this approach for all master prospects and Indian prospects are six. Therefore, after knowing the suitable number of clusters, non-hierarchical variation of clustering for categorical/binary data, i.e. K-modes can be conducted and the result of the analysis has been described in the following section.

4.3. Evaluation Phase (Cluster Analysis)

As mentioned in Chapter 3, explanatory variable for this study is application dossier submitted via Osiris (student information system), which is one of the pivotal conversion stages for Marketing and communication department of University of Twente. Moreover, departments acknowledge that before this final conversion (submission of application dossier), website visitors interested in master studies leaves behind digital trails of behaviour they manifest on the University website and unravelling these behavioural patterns is necessary to enhance the effectiveness of marketing effort. Further, in this study, considering time frame corresponding to each event hit (behavioural attribute) is doubtable because usually, every visitor represents a different sequence for manifesting behavioural features, for instance, visitor 1 first downloaded PDF, then requested an education brochure, on the other hand, visitor 2 first requested for education brochure and then, downloaded PDF. Therefore, in both cases, visitors follow different sequences but share similar behaviours. Further, if both visitors are converted (submitted application) then analysing similarity between behavioural attributes is more beneficial than the analysis based on a specific sequence they follow. Furthermore, this section presents the fifth phase of the CRISP-DM model, i.e. evaluation phase in which clustering patterns are analysed, and cluster must be evaluated for quality and effectiveness before their use in the field.

4.4.1 Behavioural profiling of All Master visitors

After identifying the number of clusters for All Master visitors via hierarchical clustering, the output is given as input to K-modes algorithm, which distributes All Master visitors (N =2327) in six clusters. Table 5 denotes that Cluster 1 has the majority of visitors, i.e. 58.5%, followed by Cluster 2 which has 21.1% of visitors, which implies that these two clusters have roughly 80% of the visitors and the rest 20% is distributed in Cluster 3 (4.6%), Cluster 4 (5.5%), Cluster 5 (3.6%) and Cluster 6 (6.6%).

Table 5: Distribution of All Master visitors in each cluster

Distribution of All Masters visitor in each Cluster			
	N	*%*	*Valid %*
Cluster 1	1362	58.5	58.5
Cluster 2	492	21.1	21.1
Cluster 3	108	4.6	4.6
Cluster 4	127	5.5	5.5
Cluster 5	84	3.6	3.6
Cluster 6	154	6.6	6.6
Total	2327		

In Cluster 1, merely 6.9% (refer table 6) of the visitors are converted, and prominent behavioural attribute of these visitors are Managed CTA (Call to Action) and Education Brochure Request. In marketing CTA is instruction given to the audience to evoke an immediate response for instance 'Register now, 'Apply now' etc. In Managed CTA, when the user revisits the website, content and message of the CTA is automatically tailored to persuade the visitor to act. Managed CTA click happens when a visitor clicks the CTA button. Roughly, 92% of the visitor in Cluster 1, requested for education brochure, nearly 2 in 5 clicked the Managed CTA, and roughly 1 in 10 downloaded PDF, searched for scholarship and requested for to become a student for a day. Rest of the behavioural attributes are negligible. Mostly, visitors downloaded the Education brochure; it implies that these brochures are the main source of information for Cluster 1. Therefore, we can say that they are in the awareness phase, gathering knowledge about the university. Quite a few visitors searched for scholarships, FAQs or requested student for a day that is why conversion rate of this cluster is quite less. Hence, this Cluster can be labelled as *'Moderately aware Prospects' or 'MA'.* Further, 64% of the visitors landed on the university website through Google, 10% directly and the rest of the visitors through the remaining 129 behavioural source features. Among all visitors in Cluster 1 majority (70.3%) prefers to surf the webpages via desktop. Table 9 represents the top 3 countries from each continent in term of traffic of visitors. Dutch visitors constitute 35.2% of the total traffic, followed by Indians (13.7%) and Indonesians (5.7%). Also, a most popular study in this group is Sustainable Energy Technology.

In Cluster 2, 18.7% submitted the application. In this cluster all visitors went through eligibility check, 40.2% clicked managed CTA, nearly 1 in 5 of downloaded the PDF and searched for scholarship and the remaining attributes are almost trivial. Eligibility-Check is a dominant attribute as it informs the visitors whether they meet the minimum requirements for a specific study program they are interested in. Therefore, we can say that visitors belonging to this cluster are interested in the University of Twente for their future studies. However, a low proportion of visitors read any sort of information material available on the website, to know about the university. Likely reasons, which creates interest in the visitors to study at the University of Twente are the positive spread of WOMs, offline and online advertising, and e-WOMs via disparate social media channels. Hence, this cluster can be labelled as *'Interested' (Interested prospects).* Moreover, lion share's (58.7%) of the visitor enters the website via Google. Also, 78.5% prefer desktop to explore the website. Table 9 illustrates that national visitor forms 20.9% of the total proportion and a maximum number of internationals are from India (5.3%), Indonesia (5.3%), Germany (4.1%) and Nigeria (3.25%). Lastly, a well-liked study in this group is Business Administration.

Conversion of the visitor in Cluster-3 is 74.1%, therefore, its worth to consider what kind of behaviour are prominent in this cluster. In this cluster, almost every visitor went through Eligibility-Check, which implies they are interested in the University courses. However, a few number of people downloaded the PDF or requested for Education Brochure. Nevertheless, 9 in 10 visitors, at least once went to frequently asked questions webpage, which raises their awareness about the university. Also, every visitor asked a question via web form which implies visitors needs to ask about specific information pertaining to their interest which is not available on the university website, or they accidentally skipped the information or not bothered to read the entire information available on the website. The relevant replies (with relevant content) to the questions of visitors, raise awareness and create interest and desirability. Desirability can be captured through, registration for open day and request to become a student for a day by visitors. The high conversion rate in cluster signifies the importance of every single stage or touch point in the customer journey is necessary to build the customer experience, which leads to high conversion. Therefore, this cluster can be labelled as *'Moderately aware High Potential prospects' or 'MA-HP*'. Mostly, visitors enter the website through Google (55.6%) and navigate the web pages from the desktop (81.5%). Majority of the internationals in this cluster are from India (10.25%), Nigeria (4.63%) and Pakistan (4.6%). Furthermore, a well-liked study in this group is Environmental and Engineering Management (22.2%).

Cluster 4 has the highest conversion among all clusters, which is 83.5%. Out of nine behavioural attributes considered for this study, at least 8 in 10 visitors, manifested eight behavioural features, and every visitor went through eligibly check, asked a question via a web form, clicked managed CTA and requested for education brochure. However, quite a few went through the eligibility check. Visitors in this cluster are highly aware of the University in disparate aspects. Further, their curiosity/interest in the university can be seen as every visitor asked a question via web form to explore beyond the information which is available on the university website. Key difference w.r.t Cluster 3 is that visitors are moderately aware of the university in cluster three, but in cluster four, they are highly aware of the university. Further, their keen desire to experience the study and university culture indicates that most of the stages or touch points pertaining to their customer journey went quite well which is why conversion for this cluster is highest (83.5%) which is nearly 12.1 times the Cluster 1, which has the majority (58.5%) of the visitors. Hence, these can be labelled as *'Desired high potential prospects' or 'Desired-HP'*. In this cluster, there are two prominent sources to enter the website, i.e. Google (30.7%) and quicklink (26%). There are roughly 30% Dutch visitors in this cluster and among internationals Indians (15.7%), and Germans (6.3%) are in the majority. A most famous study among this cluster is Sustainable Energy Technology.

Cluster 5 has the second highest conversion that is tantamount to 78.6%. It is quite similar to cluster 4, i.e. even in this cluster nearly 8 in 10 visitors manifested eight behavioural attributes, and every visitor went through eligibility check, requested to become a student for a day and education brochure. However, not even a single visitor asked a question via the web form which implies whether visitors are either satisfied with the available information or reluctant to ask for more information via web form. The latter reason seems to valid as conversion in this cluster is roughly 5% less than the Cluster 4. Hence, it can be labelled as *'Reluctant high potential prospects' or 'Reluctant-HP'*. Again, in this cluster majority enters the website through Google (61.9%) and nearly 90% prefer to navigate website via desktop. Moreover, most numbers of Internationals are from India (10.7%), Indonesia (7.1%), Brazil (4.8%) and Ecuador (4.8%). Lastly, the well-liked study is business administration.

In cluster 6, 14.3% of the visitors are converted. There are two prominent behaviours in this cluster, i.e. Managed CTA click and Registration for the open day; both have an identical percentage of visitors which is 91.6%. Request to become a student for a day and eligibility check is moderate among visitors. Merely, 2 in 5 visitors downloaded education PDF and requested an education brochure, and the rest of the behavioural attributes is nearly negligible. Therefore, the majority of visitors are in the desirability stage, i.e. they have the desire to experience the student life at the University of Twente on Open days. However, conversion is low which illustrates that how vital it is to build a comprehensive experience, by making the customers pass through every possible touch point but it should be relevant. Since experience drives prospects in this cluster; it can be labelled as *'Trial driven prospects' or 'Trail driven'*. 76.6% considered Google to visit university website and 85.1% surf the website through the desktop. Three-fourth of the visitors are from the Netherlands after that maximum visitors are from Germany. A most popular study in this cluster is Business Administration.

Table 6: Distribution of behavioural attributes of All Master visitors in each cluster

Distribution of Behavioural Attributes of All Master visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	MA		Intersted		MA-HP		Desired-HP		Reluctant-HP		Trial driven	
Behavioural Attributes	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Osiris Application Submitted	94	6.9	92	18.7	80	74.1	106	83.5	66	78.6	22	14.3
Managed CTA Click	271	19.9	198	40.2	89	82.4	127	100.0	67	79.8	141	91.6
PDF Download	135	9.9	97	19.7	21	19.4	112	88.2	67	79.8	25	16.2
Scholarship Finder	100	7.3	70	14.2	43	39.8	123	96.9	73	86.9	11	7.1
Education Broucher Request	1249	91.7	0	0.0	5	4.6	127	100.0	84	100.0	31	20.1
Registration Open Day	12	0.9	23	4.7	106	98.1	116	91.3	80	95.2	141	91.6
FAQs	37	2.7	12	2.4	100	92.6	126	99.2	78	92.9	6	3.9
Question via Webform	0	0.0	0	0.0	108	100.0	127	100.0	0	0.0	12	7.8
Request student for the day	113	8.3	0	0.0	105	97.2	107	84.3	84	100.0	64	41.6
Eligibility Check	0	0.0	492	100.0	106	98.1	20	15.7	84	100.0	47	30.5

Table 7: Distribution of well-liked studies in each cluster

Distribution of well liked studies in each cluster

Cluster-1			Cluster-2			Cluster-3			Cluster-4			Cluster-5			Cluster-6		
MA			Intersted			MA-HP			Desired-HP			Reluctant-HP			Trial driven		
Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*
SET	89	6.5	BA	44	8.9	EEM	24	22.2	SET	17	13.4	BA	24	28.6	BA	29	18.8
IEM	87	6.4	SET	42	8.5	SET	10	9.3	BIT	10	7.9	SET	7	8.3	IEM	15	9.7
CEM	81	5.9	CEM	40	8.1	SE	8	7.4	ME	9	7.1	CEM	7	8.3	ME	11	7.1
EEM	77	5.7	ME	36	7.3	MRM	8	7.4	BA	9	7.1	EE	6	7.1	EST	8	5.2
SE	74	5.4	CS	34	6.9	CEM	6	5.6	P	8	6.3	HS	5	6.0	CEM	7	4.5
BA	71	5.2	IEM	26	5.3				IEM	7	5.5	CS	4	4.8	HS	7	4.5
HS	70	5.1	BIT	26	5.3										TM	7	4.5
ME	69	5.1	HS	24	4.9										SET	6	3.9
CEM	61	4.5	CME	22	4.5												
MSM	59	4.3	CE	21	4.3												

Table 8: Distribution of All Master visitors source attribute in each cluster

Distrbution of All Master visitor Source Attributes in each cluster

Cluster-1			Cluster-2			Cluster-3			Cluster-4			Cluster-5			Cluster-6		
MA			Intersted			MA-HP			Desired-HP			Reluctant-HP			Trial driven		
Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*
google	871	64.0	google	289	58.7	google	60	55.6	google	39	30.7	google	52	61.9	google	118	76.6
(direct)	136	10.0	quicklink	36	7.3	(direct)	13	12.0	quicklink	33	26.0	(direct)	4	4.8	(direct)	11	7.1
mastersportal	42	3.1	(direct)	35	7.1	quicklink	9	8.3	(direct)	15	11.8	quicklink	3	3.6	mail.google	3	1.9
bing	40	2.9	mastersportal	22	4.5	mastersportal	3	2.8	outlook	11	8.7	mail.google	3	3.6	mail-msc	3	1.9
quicklink	39	2.9	outlook	18	3.7	bing	3	2.8	utwente.nl	4	3.1	outlook	2	2.4	quicklink	2	1.3

Table 9: Distribution of All Master visitors in each cluster continent wise

Distribution of All Master visitor country in each cluster continent wise

Cluster-1			Cluster-2			Cluster-3			Cluster-4			Cluster-5			Cluster-6		
MA			Intersted			MA-HP			Desired-HP			Reluctant-HP			Trial driven		
Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*	Country	*N*	*%*
Europe																	
Netherlands	479	35.2	Netherlands	103	20.9	Netherlands	24	22.2	Netherlands	38	29.9	Netherlands	22	26.2	Netherlands	118	76.6
Germany	35	2.6	Germany	20	4.1	Germany	3	2.8	Germany	8	6.3	Greece	4	4.8	Germany	14	9.1
Italy	23	1.7	Spain	10	2.0	Italy	2	1.9	Albania	3	2.4	Germany	3	3.6			
Asia																	
India	186	13.7	India	67	13.6	India	11	10.2	India	20	15.7	India	9	10.7	India	3	1.9
Indonesia	77	5.7	Indonesia	26	5.3	Pakistan	5	4.6	Indonesia	6	4.7	Indonesia	6	7.1	Indonesia	1	0.6
Iran	23	1.7	Pakistan	12	2.4	Indonesia	4	3.7	Pakistan	4	3.1	Iran	3	3.6			
North America																	
Mexico	19	1.4	Mexico	13	2.6	Mexico	4	3.7	Mexico	3	2.4	Mexico	2	2.4			
USA	15	1.1	USA	3	0.6				USA	2	1.6						
Canada	3	0.2	Canada	3	0.6						0.0						
South America																	
Brazil	25	1.8	Brazil	17	3.5	Brazil	4	3.7	Brazil	2	1.6	Brazil	4	4.8	Chile	1	0.6
Colombia	14	1.0	Colombia	5	1.0	Colombia	2	1.9				Ecuador	4	4.8			
Ecuador	7	0.5	Suriname	2	0.4	Bolivia	1	0.9									
Africa																	
Nigeria	49	3.6	Nigeria	16	3.25	Nigeria	5	4.63	Ghana	5	3.94	Ghana	3	3.57	Uganda	2	1.3
Ghana	31	2.28	Ghana	14	2.85	Uganda	4	3.7	Nigeria	3	2.36	Nigeria	3	3.57	Nigeria	1	0.65
Kenya	19	1.4	Kenya	8	1.63	Kenya	2	1.85	Uganda	2	1.57	Egypt	2	2.38			
Oceania																	
Australia	5	0.37	Australia	2	0.41	New Zealand	1	0.93	Australia	2	1.57	Australia	1	1.19			

Table 10: Distribution of preferred device type of All Master visitors in each cluster

Distribution of preferred device type by All Master visitor in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	MA		Intersted		MA-HP		Desired-HP		Reluctant-HP		Trial driven	
Device type	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
desktop	958	70.3	386	78.5	88	81.5	102	80.3	75	89.3	131	85.1
mobile	349	25.6	94	19.1	15	13.9	24	18.9	9	10.7	21	13.6
tablet	55	4.0	12	2.4	5	4.6	1	0.8	0	0.0	2	1.3

4.4.2. Behavioural profiling of Indian Master Visitors

As mentioned in the prior chapters, an optimal number of clusters suggested in Stage 1 algorithm for Indian Master prospects is six, therefore, in Stage 2, i.e. K-modes algorithm distributes all visitors (N=296) in six clusters. Cluster 4 has the highest number of visitors, i.e. 106 (35%) whereas Cluster 6 has the smallest, which is tantamount to 3% of the total visitors. Cluster 4 and Cluster 6 constitutes 2/3 of the visitors whereas rest of the 1/3 visitors are distributed among Cluster 1 (13.9%), Cluster 2 (12.5%), Cluster 3 (3%) and Cluster 5 (6.8%).

Table 11: Distribution of Indian Master visitors in each cluster

Distribution of Indian (Master) visitors in each Cluster

	N	*%*	*Valid %*
Cluster 1	41	13.9	13.9
Cluster 2	37	12.5	12.5
Cluster 3	9	3.0	3.0
Cluster 4	106	35.8	35.8
Cluster 5	20	6.8	6.8
Cluster 6	83	28.0	28.0
Total	296		

Table 12 represents the distribution of behavioural attributes of Indian Master visitors in each cluster. In comparison to All Master prospects, one behavioural attribute is entirely missing for Indian master visitors, i.e. registration for an open day. It is expected because it is highly inconvenient in terms of the VISA process and financial expenses just to experience student life for one day. Although roughly, one in 40 visitors in Cluster 1, Cluster 2 and Cluster 3 are registered for the open day, even this is an insignificant number. Therefore, we can say PDF download, educational brochure, scholarship finder, FAQs, a question via web form and eligibility check are the remaining sources for information and interaction for visitors.

Cluster 1 has the conversion rate of 17.1%. Distinguished behavioural feature in this cluster is scholarship finder and education brochure. Every member of this cluster at least once searched for scholarship on university webpage and 85.4% requested for education brochure. Nearly, 20% downloaded PDF, 1 in 6 went through the eligibility check, and the rest of all behavioural attributes are trivial because rarely any visitor manifested that behavioural features. High request for education brochures and low for PDF downloads indicates that these visitors refer educational brochure as their main source of information about the university and unaware about the rest of the sources or ignored to know more about the university. However, scholarship appears to be a pivotal factor in considering education at the University of Twente for this group of visitors. Therefore, this group can be labelled as *'Scholarship driven prospects'*.

Cluster 2, Cluster 4 and Cluster 5 have similar distribution of behavioural attributes, but conversion rates vary from 6.6% to 20%. All the visitors in these groups at least once requested for education brochure, approximately 10 to 20% visitors downloaded PDF. Table 12 figures lucidly illustrate the variation in the conversion w.r.t to PDF downloads. In Cluster 4, 6.6% visitors are converted corresponding to 12.3% PDF downloads; cluster 2 has 16.2% conversion rate in which approximately 19% visitors downloaded the PDF and Cluster 5 has converted 20% of the visitors in which collectively 20% of the visitors searched for scholarship

and downloaded the PDF. In addition, in cluster 2 and cluster 5, the number of visitors who at least once visited the frequently asked questions webpage is twice as compared to Cluster 4. Therefore, it is suitable to label Cluster 2 and cluster 5 as moderately aware prospects. Therefore, the Cluster 2 can be tagged as *'MA-1a'* and Cluster-5 as *'MA-1b'*. Also, Cluster 4 has partially to moderately aware prospects. Hence it can be tagged as *'MA-2'*.

Cluster 3 has the highest conversion rate, i.e. 44%, in which every visitor asked a question via a web form which implies that the visitors in this cohort are curious to know more about university or searching for information which is not available in PDFs. Therefore, it can be hypothesized that this group of visitors are interested in exploring more about the university to support their decision regarding future studies. Hence, this group can be labelled as *'Interested high potential prospects' or 'Interested-HP'*.

In Cluster 6, the number of visitors converted is 13.3%, and every visitor went through Eligibility-check. This behavioural attribute often manifested by visitors who are curious to know whether they meet the minimum requirements for admission into the university, i.e. they are interested in the university studies. Table 12 represents that only 10 or less in every 100 visitors have downloaded PDF or educational brochures, which implies they are partially unaware or do not have comprehensive knowledge about University. Therefore, it can be labelled as *'Interested' or 'Interested prospects'*.

Table 12: Distribution of behavioural attribute of Indian Master visitors in each cluster

Distribution of Behavioural Attributes of Indian (Master) visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Scholarship driven		MA-1a		Interested-HP		MA-2		MA-1b		Intersted	
Behavioural Attributes	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Osiris Application Submitted	7	17.1	6	16.2	4	44.4	7	6.6	4	20	11	13.3
Managed CTA Click	27	65.9	37	100.0	1	11.1	0	0.0	0	0	21	25.3
PDF download	8	19.5	7	18.9	3	33.3	13	12.3	2	10	13	15.7
Scholarship finder	41	100.0	0	0.0	0	0.0	0	0.0	2	10	7	8.4
Education broucher request	35	85.4	37	100.0	0	0.0	106	100.0	20	100	0	0.0
Registration Open Day	1	2.4	1	2.7	0	0.0	0	0.0	0	0	2	2.4
FAQs	2	4.9	2	5.4	1	11.1	3	2.8	1	5	1	1.2
Question via Webform	0	0.0	0	0.0	9	100.0	0	0.0	0	0	0	0.0
Eligibility Check	6	14.6	0	0.0	0	0.0	0	0.0	0	0	83	100.0

Table 14: Distribution of well-liked studies in each cluster

Distribution of well liked studies in each cluster

Cluster-1			Cluster-2			Cluster-3			Cluster-4			Cluster-5			Cluster-6		
Scholarship driven			MA-1a			Interested-HP			MA-2			MA-1b			Intersted		
Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*	Studies	*N*	*%*
SET	5	12.2	IEM	6	16.2	BA	4	44.4	ME	16	15.1	IEM	3.0	15.0	ME	18	21.7
ME	5	12.2	SET	4	10.8	SET	2	22.2	IEM	13	12.3	GSEO	3	15.0	IEM	12	14.5
IDE	5	12.2	ME	4	10.8	N	1	11.1	CE	10	9.4	ME	2	10.0	CE	9	10.8
IEM	3	7.3	ES	3	8.1				EE	7	6.6	CM	2	10.0	SET	8	9.6
BA	2	4.9	IDE	2	5.4				ES	7	6.6	SET	2	10.0	IDE	4	4.8
CE	2	4.9	CE	2	5.4				SE	7	6.6				EE	3	3.6
CEM	2	4.9	CEM	2	5.4				SET	5	4.7				BA	3	3.6

Table 16: Distribution of preferred device type of Indian Master visitors in each cluster

Distribution of preferred device type by Indian (Master) visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Scholarship driven		MA-1a		Interested-HP		MA-2		MA-1b		Intersted	
Device type	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
desktop	21	51.2	25	67.6	9	100.0	57	53.8	14	70.0	57	68.7
mobile	20	48.8	12	32.4	0		46	43.4	6	30.0	26	31.3
tablet	0	0.0	0		0		3	2.8	0	0.0	0	0.0

4.5. Interpretation of Analyses

In prior sections of this chapter, six behavioural profiles were discovered among all website visitors who are keen in the master studies. These profiles have been labelled as moderately aware prospects, interested prospects, moderately aware high potential prospects, desired high potential prospects, reluctant high potential prospects and trial driven prospects. Whereas, six behavioural profiles discovered for Indian master visitors represents significant differences in behavioural attributes in comparison to profiles revealed for all master visitors. Indian behavioural profiles are scholarship driven prospects, moderately aware prospects-1a, interested high potential prospects, moderately aware prospects-2, moderately aware prospects-1b and interested prospects. Merely, concentrating on dominant behavioural attributes may not depict the realistic picture of profiles. Purpose of this section is to analyse insights of the previous section comprehensively to unveil the details, which can only be seen when patterns or insights are compared in a side-by-side fashion. This section initiates with a comparison of visitor's distribution, which is accompanied by a comparison of behavioural features of profiles and finally, graphically represents the variation in the distribution of studies, device type and source (the origin of traffic of the University website), e.g. google.com across each behavioural profile. Further, some sort of paradigm is required which can be incorporated with the behavioural profiles and renders direction to formulate marketing communication strategies. AIDA is such a model, which stands for Attention, Interest, Desire and Action. This model is pervasively used in marketing and advertising to designate the stages that occur from a point when the consumer initially becomes aware of a brand or product to the stage when the consumer makes a purchase decision or tries a product. In this study, the AIDA model is considered in the context of higher education industry (online presence). In the Attention stage, the visitor becomes aware of the university through various information material available on the university website. Then, visitors who become interested by learning about the university, often demonstrate behaviours like Eligibility-check. Desire is a stage in which visitors actively consider the studies of the university. Such intentions can be measured from behaviours for instance registration for an open day or request to become a student for a day. Strong developed this model in 1925. From the initial model, it had undergone numerous variations, but still, after 93 years, many researchers and academics are adopting it. Since, today information technology creates numerous online social media platforms which alter the way people socialize, communicate and influence the behaviour of the consumer (Wijaya; 2015). Wijaya, 2015 did the state-of-the-art development in the AIDA model, which is AISDALSLove model. However, this study is focused on the behavioural attributes manifested by visitors on the university website and these features do not include elements of social media. Therefore, AIDA paradigm is more suitable for this study. This model is applied to recognize the stage of behavioural profiles in this paradigm and this fulfilled by observing a naturally occurring

pattern in behavioural profiles so that it will support the marketing department to formulate marketing strategies accordingly.

Figure 9 & 10 represents the distribution of website visitors for All Master and Indian Master visitors in six behavioural profiles respectively. If the distribution is analysed in a generalised form (macro level), then it can be concluded by observing the bar graphs that the proportion of moderately aware visitors (59%) in All Mater category is roughly equal to the proportion of visitors who are moderately aware in Indian Master category, i.e. 55%. Moreover, moderately aware visitors in Indian Master category are further sub-divided into MA-1a, MA-1b and MA-2 because there are variations in the micro-features (less influential behavioural attribute), thus understanding this minute difference helps to refine the relevance of advertisement by ameliorating the targeting parameters. However, the percentage of visitors in 'Interested' profile of Indian visitor is roughly 7% more as compared to the All Master category of visitors. The further key observable difference is that there are merely 3% high potential prospects from Indian Master group whereas it is nearly 4.5 time in All Master group, which is due to the presence of national (Dutch) students. In this study, high potential prospects are all those visitors, which belong to the cluster whose conversion rate is higher than average conversion rate of all clusters. Lastly, both groups have one distinct behavioural profile, i.e. Trail-driven in All Master category and Scholarship driven in Indian Master category.

Figure 9: Distribution of All Master visitors in six behavioural profiles

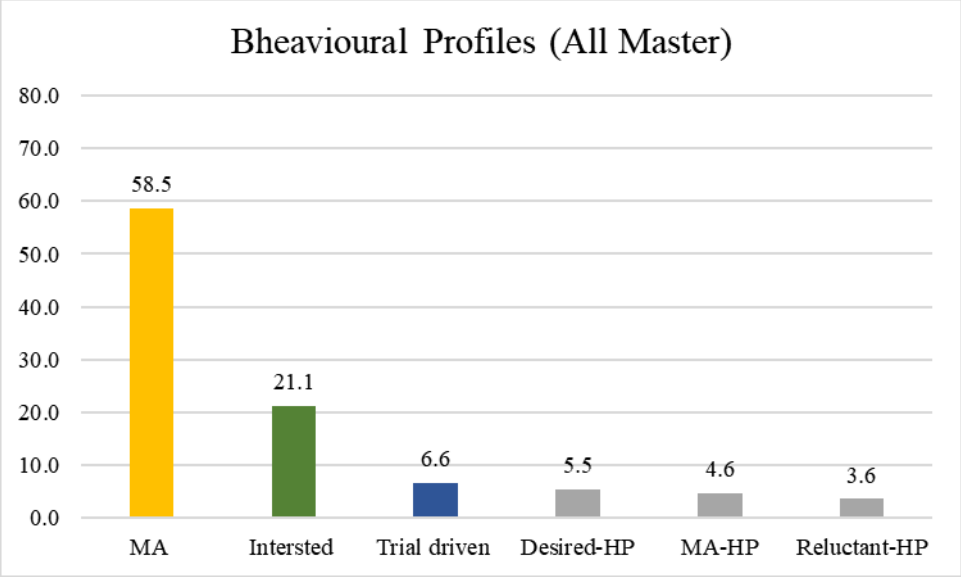
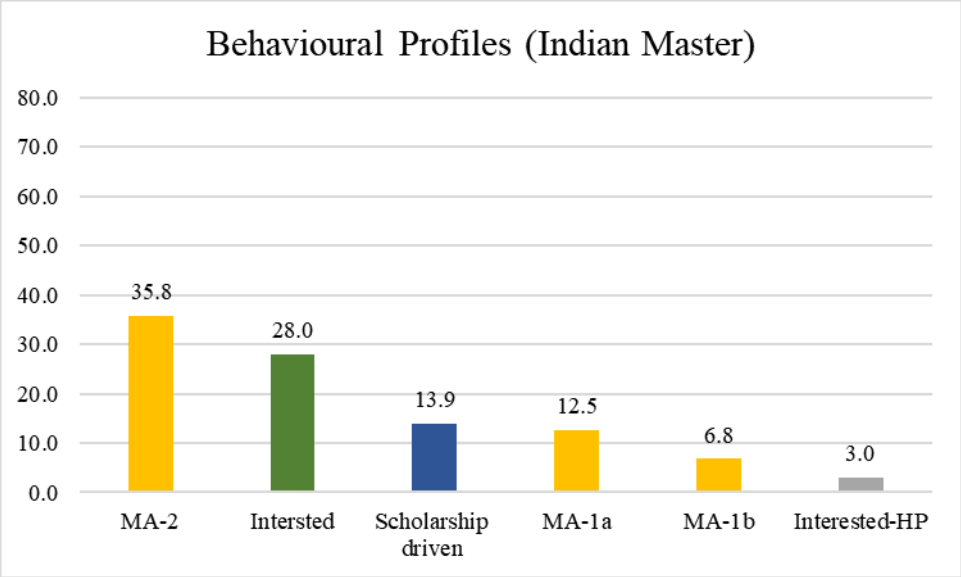


Figure 10: Distribution of Indian Master visitors in six behavioural profiles



In brief, the proportion of moderately aware visitor is alike in All Master and Indian Master group and in the rest of behavioural profiles either proportion is distinct or behavioural patterns.

Figure 11 and 12, clearly depicts the behavioural patterns that pertaining to each cluster in in All Master and Indian Master group. In the initial phase, to make interpretations thoughtful yet simple we assumed if more than 50% of the visitors pertaining to specific behavioural profile manifest certain behaviour they are considered as dominant behaviour (macro) and if the percentage of visitors are 50% or less than it is considered as a non-dominant feature. There are colossal differences in the behavioural patterns between some of the behavioural profiles of All Masters and Indian Masters. In All Master group, behavioural profiles 'MA-HP', 'Reluctant-HP' and 'Trail driven' have distinct patterns among themselves also there is no behavioural profile in Indian Master group which exhibits such patterns. Interested high potential prospects 'Interested-HP' have the highest conversion rate in the Indian master's group whereas in All Master group 'Desired-HP' has the highest percentage of converted visitors. Dominant behaviour in 'Interested-HP' profile of Indian Master group is question asked via web form, which is often manifested by visitors who are curious to know more about university due to disparate reasons some of which is already stated in section 2.2, other than that FAQs, managed CTA click, and PDF downloads are micro-features (Minutely influential attributes). Therefore, it can be hypothesized that for 'Interested-HP' profile question via web form has the strong influence on the conversion rate. On the contrary, in the 'Desired-HP' profile of All Master visitors, all attributes are dominant expect Eligibility-check.

Figure 11: Distribution of behavioural attributes of All Master visitors

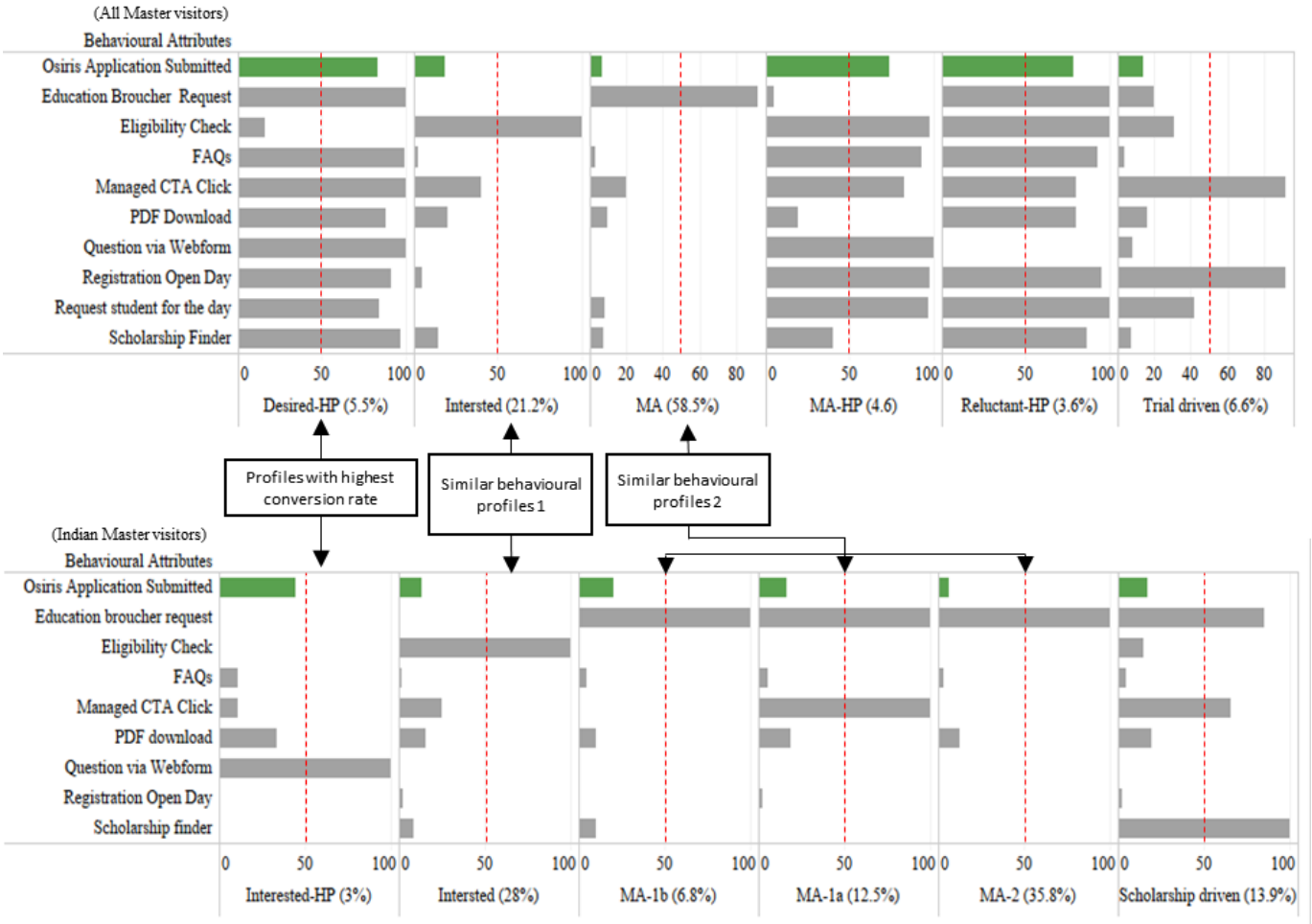


Figure 12: Distribution of behavioural attributes of Indian Master visitors

It is intriguing to observe that micro (non-dominant) and macro (dominant) behavioural features are similar in 'Interested' behavioural profile of All Master and Indian Master group which implies that conversion rate varies in a similar fashion as their macro and micro behaviour alters. It can be hypothesized that patterns in inserted visitor profile remain almost homogenous across countries, at least for India. As mentioned earlier, a moderately aware visitor in Indian Master group is sub-divided into three categories MA-1a, MA-1b and MA-2. It can be observed from Figure 12 that the macro features of these profiles is Education brochure request, and micro-features of users are FAQs, PDF download and visitors at least once visited the scholarship webpage. MA-1b has the highest conversion, followed by MA-1a and then MA-2. In MA-1a nearly 100% visitors clicked the managed CTA, which is entirely negligible in MA-1b and MA-2, although these profiles partially have the same influence on the PDF downloads and FAQs as MA-1a. Therefore, it can be hypothesized that Managed CTA click has a weak influence on the conversion rate for moderately aware users. The influence of the micro-feature 'PDF download' on MA-1b is approximately half of MA-1a and MA-2, but still, it has the highest conversion. Presence of 'Scholarship finder' in MA-1b is the self-explanatory reason behind the situation. Therefore, it can be hypothesized that the Scholarship finder has a strong influence on the conversion for the moderately aware Indian visitor interested in masters.

Figure 13 and 14 represent the behavioural source attributes of All Master group and Indian Master group respectively. It depicts at a macro-level that all behavioural profiles have the maximum number of visitors who entered the university website through Google, some visit directly and few through quicklink. However, at the micro-level the scenario is different, for example, 'Desired-HP' profile of All Master group had a proportion of visitors from Google, (direct) and quicklink as 3: 2.5: 1 respectively. However, there are no visitors from quicklink for 'Interested-HP' profile of Indian Master group and the proportion of visitors from Google and who directly visited the website is 8: 1 respectively. A similar difference is prevalent in all profiles of both groups. It can be hypothesized that at macro-level (ordinal) behavioural source attributes are similar but at micro-level (metric) it varies between All Master and Indian Master visitors.

Figure 13: Behavioural source attribute of All Master visitors

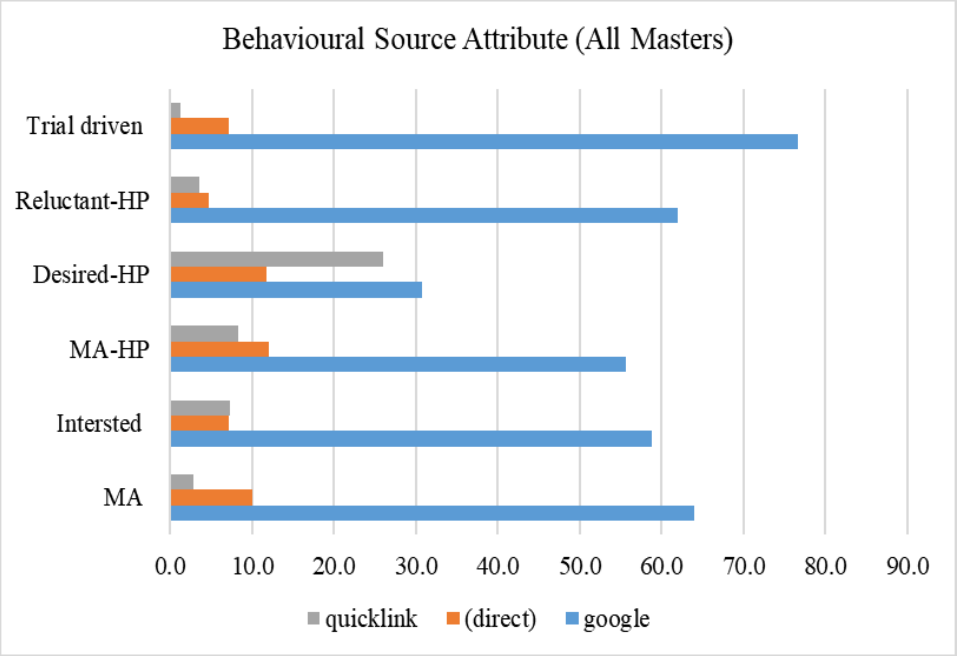


Figure 14: Behavioural source attribute of Indian Master visitors

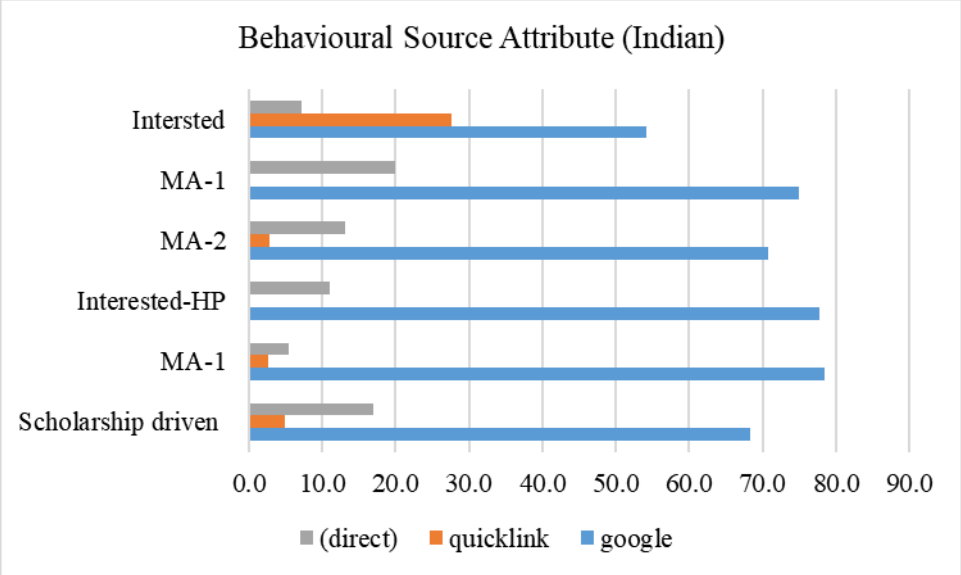


Table 16 illustrates the distribution of well like studies for All Master and Indian Master categories. Again, at macro level (ordinal) most famous studies in both groups are alike for instance Sustainable Energy Technology (SET), Mechanical Engineering (ME), Industrial Engineering and Management (IEM), Business Administration (BA) and Chemical Engineering (CE). However, at micro-level (metric) scenarios are distinct. A most famous study in 'Interested' profile of All master group is Business administration (BA), and the percentage of visitors interested are 8.9 %, whereas the famous study in 'Interested' profile of Indian Master group is Mechanical Engineering (21.7%). Therefore, it can be hypothesized that at macro-level (ordinal) interest in master studies of website visitors is similar but at micro-level (metric) it varies between All Master and Indian Master visitors.

Table 16: Distribution of well-liked studies for All Masters and Indian Master visitors

Distribution of well liked studies in each cluster (All Master vsitors)

MA		Intersted		MA-HP		Desired-HP		Reluctant-HP		Trial driven	
Studies	%	Studies	%	Studies	%	Studies	%	Studies	%	Studies	%
SET	6.5	BA	8.9	EEM	22.2	SET	13.4	BA	28.6	BA	18.8
IEM	6.4	SET	8.5	SET	9.3	BIT	7.9	SET	8.3	IEM	9.7
CEM	5.9	CEM	8.1	SE	7.4	ME	7.1	CEM	8.3	ME	7.1

Distribution of well liked studies in each cluster (Indian Master visitors)

Scholarship		MA-1		Interested-HP		MA-2		MA-1		Intersted	
Studies	%	Studies	%	Studies	%	Studies	%	Studies	%	Studies	%
SET	12.2	IEM	16.2	BA	44.4	ME	15.1	IEM	15.0	ME	21.7
ME	12.2	SET	10.8	SET	22.2	IEM	12.3	GSEO	15.0	IEM	14.5
IDE	12.2	ME	10.8	N	11.1	CE	9.4	ME	10.0	CE	10.8

Figure 15 and 16 depicts the type of device preferred by visitors to surf the University of Twente website. At the global level most, preferred device by visitors is a desktop, the second preferred device is a mobile phone and the least preferred device is a tablet. This pattern is similar in both groups. At a local (micro) level, there are significant differences, tablets are rarely used in both groups but the ratio of desktop users to the mobile phone users varies from 1:1 to 2.5:1 in Indian Master group whereas this ratio varies from 2.5:1 to 9:1 in All Master group. Hence, we can conclude that usage of mobile for surfing website by Indian Master visitors is high as compare All Mater visitor. Also, it can be hypothesized that at macro-level (ordinal) preferred device type to surf UT website by visitors interested in master studies is similar but at micro-level (metric) it varies between All Master and Indian Master visitors.

Figure 15: Preferred device to surf UT website by All Master visitors

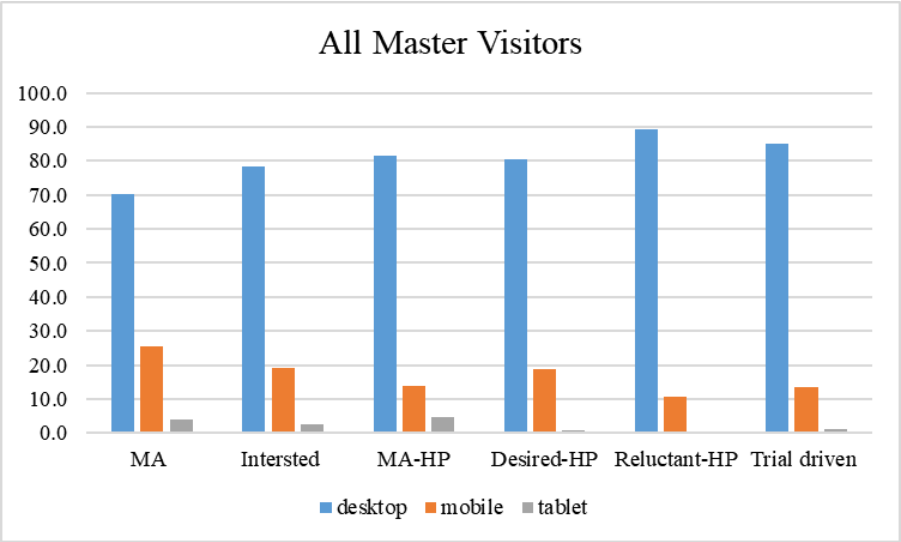
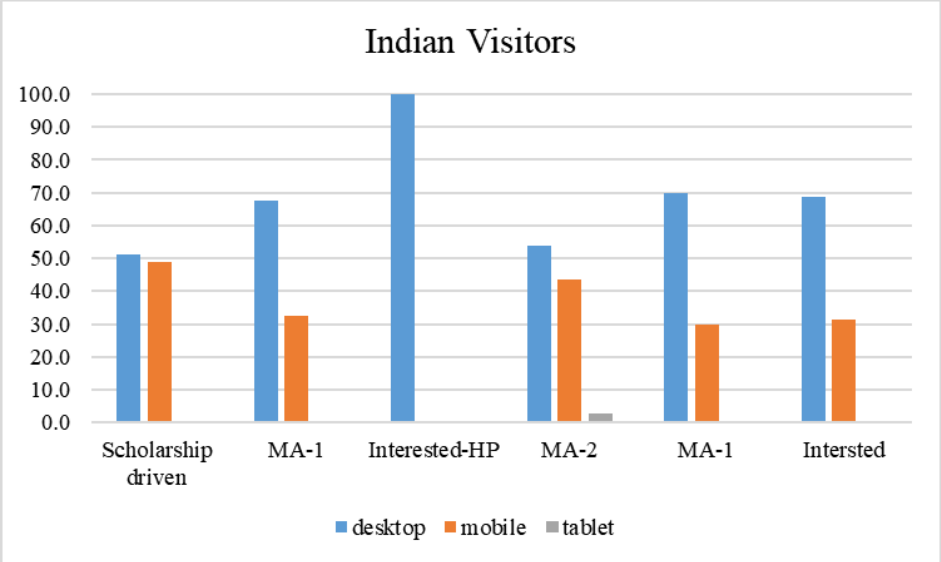


Figure 16: Preferred device to surf UT website by Indian Master visitors



Overall, there are significant differences between the All Master visitors and Indian Master visitors. Hence, it proves if we create subsets of All master visitors country wise, then behavioural patterns in a particular subset (India) varies from the whole set. Further, there is substantial variation across the behavioural source attributes, well-liked studies and preference of device to surf the university website.

4.6. Clustering Validation

Clustering is the unsupervised machine learning process thus evaluation of its results is quite important. The procedure of assessing the results of the clustering algorithm is known as

cluster validity assessment. Two dimensions have been proposed for assessing and selecting an optimal clustering scheme (Halkidi et al., 2001): first is the compactness, which implies that the members pertaining to each cluster, must be close to each other as much as possible. Second is separation; it implies clusters must be widely separated.

Numerical measures, which are often applied to numerous aspects of cluster validity, are categorised into following three types (Halkidi et al., 2001): External Criteria, Internal Criteria and Relative Criteria. The external validity method assesses the clustering based on user-specific intuition which implies the degree to which the cluster label overlaps the externally supplied class labels (ground truth is available). The internal criteria is used to measure the quality of the clustering structure when the ground truth about the dataset is unavailable. The relative criteria is often used to compare two or more distinct clusters or clustering techniques. Dataset used in this study neither has the availability of ground truth, nor is it used to compare with other clustering techniques. Therefore, internal or intrinsic criteria is used. The silhouette coefficient is such a measure which is extensively used in prior researches. There are other methods as well but none of them have an advantage over the other. Silhouette analysis is often used in literature for cluster validation (Muguerza, Pérez, & Perona, 2013; Pollard & van der Laan, 2002).

4.6.1. Silhouette analysis (Internal Criteria)

The Silhouette analysis measures the extent to which the observations are well-structured, and it analyzes the average distance between clusters (Jain, 2016; De Amorim & Hennig, 2015). Silhouette score measures the unity within-cluster via distances of users in the same cluster from its centroid (De Amorim & Hennig, 2015). The scores vary from -1 to 1, where score near to -1 implies unity/cohesion within the clusters is poor and a score of 1 implies that cohesion is nearly perfect.

Still, it's challenging to select the appropriate method for executing the Silhouette analysis. For numerical/metric datasets, the structure of clustering is often validated by the density distribution and geometry of clusters. When distance function is rendered to the metric dataset, it's obvious to utilise the density-based methods into clustering (Kriegel and Sander, 1999). Further, due to the absence of intuitive distance functions between categorical variables, the methods used in the cluster validation for metric data are not relevant to categorical data (Chen et al.; 2009). Liu et al.; 2018 for Silhouette analysis used simulated binary datasets which have been partitioned by the PAM (partition around medoids) algorithm, with simple matching distance (SMD) (e.g. Gower, 2004).

Figure 17: Silhouette analysis of All Master visitors

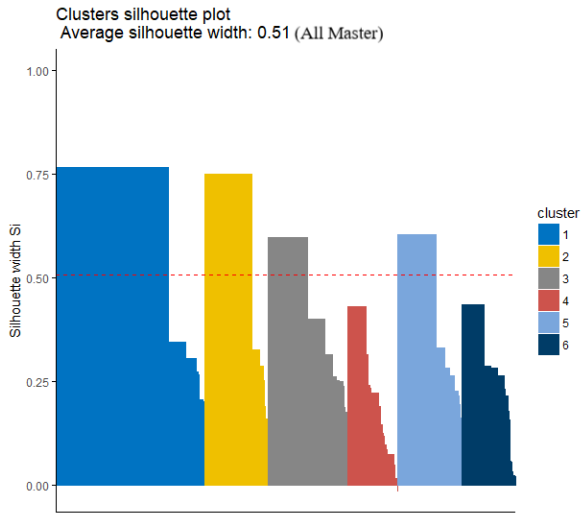
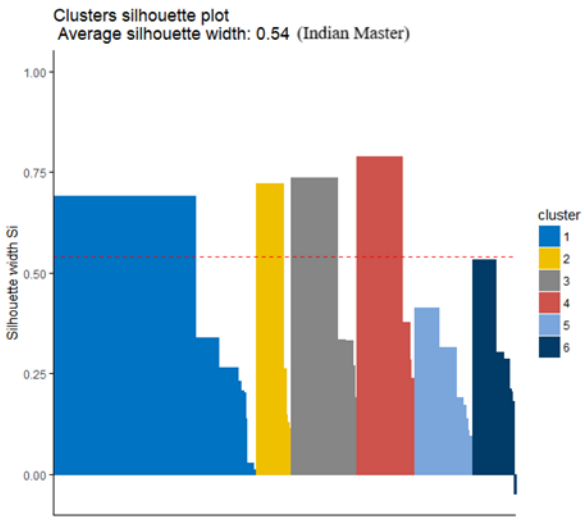


Figure 18: Silhouette analysis of Indian Master visitors



In brief, for Silhouette analysis, the PAM algorithm is adopted the results and the results have been concluded in Figure 17 and 18 (limitation it is still using distance measure, not density-measures). For both All master visitors and Indian Master visitor silhouette, width score is 0.5 and 0.55 respective, which indicates the clusters are good and reliable.

5. Discussion and Conclusions

5.1. Discussion

The objective of this study was to outline the framework for profiling of the website visitors based on the digital trails they leave behind during interaction with the website. Srimani (2011) illustrates that there are four indispensable stages of behavioural targeting. Among them, the third stage i.e. to generate a user profile using an advanced algorithm is quite critical and strenuous task. Often profiling is done via a post-hoc (data-driven) method, and unsupervised machine-learning algorithms are used for clustering/ profiling of customer attributes. Therefore, a framework is developed which will assist to recognize the suitable techniques pertaining to unsupervised machine learning for a specific dataset which can varies along the parameters volume, ability to handle dimensionality and most importantly whether the dataset is nominal/ categorical. As discussed in Chapter 2, there are numerous techniques in literature, which performs, efficiently on numeric/metric dataset but there are a few which suit the requirements of the categorical dataset. The framework for user profiling by unsupervised machine learning techniques can found in the Chapter 2 section 2.9.

Moreover, a model is presented which indicates the effect of the nature of data on the quality and interpretability of profiles. Dolnicar (2008), mentioned three approaches to user profiling: ‘a priori’, ‘a posteriori’ and ‘Hybrid’. Amongst these approaches, the ‘Hybrid’ is recommended by Dolnicar (2008), as it balances the merits and demerits of the other two approach to achieve a robust user profiling. Cufoglu (2014) stated two ways to obtain information about the user, i.e. implicit and explicit. He proposed profiling is more proficient when it is a blend of the implicit and explicit methods. The model can be found in Chapter 2 section 2.10. With the aid of the proposed framework and model, six behavioural profiles for Indian Master and All Master visitors are discovered. Clusters of All Master visitors are profiled as moderately aware prospects, interested prospects, moderately aware high potential prospects, desired high potential prospects, reluctant high potential prospects and trial driven prospects. Six behavioural profiles discovered for Indian Master visitors represents sharp differences in comparison to profiles revealed for All Master visitors. Profiles of Indian master group are scholarship driven prospects, moderately aware prospects-1a, interested high potential prospects, moderately aware prospects-2, moderately aware prospects-1b and interested prospects. At macro-level (dominant behavioural attributes), the proportion of moderately aware visitor is nearly identical in All Master, and Indian Master category and the rest of the behavioural profiles are different in term of percentage of visitors as well as the proportion of behavioural attributes. Behavioural profiles are partially distinguishable by key or dominant behavioural attributes of visitors because the percentage of key behavioural attributes is somewhat similar but non-dominant behavioural features have a unique pattern for each profile, thus making them distinct.

Furthermore, most of the research in the domain of marketing specifically behavioural targeting focused towards the metric behavioural features such as CTR (Click-through rate),

time spent on a webpage, number of pages navigated etc. For instance, Etminani et al. (2009) used the pre-processed weblogs of university website (http://www.um.ac.in) such as number of mouse clicks, time spent on web pages etc. to discover website navigation patterns of visitors from the web data to determine the factors influencing the submission of enrolment application via model-based (self-organising maps) machine learning technique. However, this research uses genuine behavioural interaction of website visitors. This was realised by utilizing multiple data-sources, namely Google Analytics and CRM database.

Also, it was concluded that concentrating on dominant behavioural attributes might not depict the realistic picture of profiles. To verify that profiles of All Masters and Indian Master visitors are compared in a side-by-side fashion to get deep insights. The outcomes reveal that dominant and non-dominant behavioural attributes vary in both categories. Besides that, at macro-level (ordinal) behavioural source attributes, well-liked studies and preferred device to surf the website are similar but at micro-level (metric) it varies between All Master and Indian Master visitors. Moreover, we hypothesized that the profile with the highest conversion rate in the Indian Master visitors category, i.e. 'Interested-HP' profile is strongly influenced by 'question via web form'. To test the hypothesis or effectiveness of behavioural profiles created for behavioural targeting, deployment of online marketing campaigns is essential. Deployment phase is the last the stage of the CRISP-DM process, which states behavioural profiling does not illustrate the completion of the project, it is crucial to make use of the created model. There are two methods to deliver targeted messages to the website visitors, first is network behavioural targeting and second is onsite behavioural targeting (Srimani, 2011).

Network behavioural targeting is the method with which majority of marketeers are familiar. The advertisement network on the web has several websites as their publishers. These publishers show advertisements on behalf of the advertisement network. Google AdWords is one such ad network developed by Google, where advertisers pay to show advertisements, product listing, video content and service offering within Google Ad network to web users. Remarketing is an AdWords feature which allows marketers to display advertisements to users that earlier visited the website. It also permits the marketer to create a distinct audience based on the behaviours of website visitors to serve the relevant advertisements to the profiled audience. For instance, it has been observed from the behavioural attributes of 'Interested-HP' profile that other than 'question via web form', visitors manifest the FAQs behaviour, therefore, designing a marketing campaign such as "Do you have any questions/How can we help you?" for Indian Master visitors who visited the FAQs webpage, can play a crucial role to enhance the conversion.

On-site behavioural targeting uses the identical concept of behavioural targeting adopted by ad networks for a customized visitor experience on the single specific website. It will succour in personalizing the webpages to improve the user experience of the visitor, ultimately driving them deeper into a website with higher engagement. Based on the learned behaviours from profiling, a marketer can set up the rules which automatically serves the content that resonates with the intent of each visitors and drive action. For instance, if Indian Master visitors (considered in the study) surf the university's FAQs webpage, and it's known from outcomes of this study, 'question via web form' have a strong influence on the conversion of 'Interested-HP' profile, then placement of call to action on FAQs as webpage displaying messages for instance "Do you have any questions" can enhance the engagement of users thus the conversion rate. Also, the integration of tools, for example, Crazy Egg can help in optimizing the placement of call-to-action or advertising (https://www.crazyegg.com/).

Lastly, the study unveils that why it is pivotal to consider the pros and cons of each unsupervised machine learning technique precisely before implementing them. Also, the necessity to treat the data at micro-level is essential, which can be seen in this study from the differences between All Master visitors and Indian Master visitors, where Indian Master visitors is a subset of All Master visitors.

5.2. Conclusion

In this research, the objective was to propose the framework for the use of unsupervised machine learning, which allows the segmentation of the disparate variety of customer attributes. By critically reviewing the literature on behavioural targeting, customer attributes, unsupervised machine learning techniques and similarly measures for the binary dataset. This study developed a framework for multi-stage clustering/ profiling of customer attributes in which distinct approaches are suggested based on the volume, ability to handle dimensionality and the type (categorical/metric) of the dataset. Along with that, a model is presented which indicates the effect of the nature of information on the quality and interpretability of clusters. Another aim was to discover behavioural profiles of Indian website visitors of the University of Twente interested in master studies therefore with the implementation of the framework and a model, six behavioural profiles are discovered. Outcome reveals that interested high potential prospects profile, i.e. 'Interested-HP' profile of Indian Master visitors has the highest conversion rate (percentage of users submitted the application) and it was hypothesized that this profile is strongly influenced by behavioural attribute 'Question via web form'. Other behavioural features, which influence Indian prospects are 'Scholarship finder' and 'Education brochure request'. Results also reveal that at macro-level (ordinal) behavioural source attribute, preferred device type and well-liked studies are similar across subsets (country wise) for All Master visitors at least for Indian website visitors interested in master studies. However, at micro-scale (metric), there are significant differences in behaviours. The aftermath of this study has numerous practical and theoretical implications, yet it subjected to many limitations, which can be used as a direction for future research.

5.2.1. Theoretical Implications

This study proposed the framework for segmentation of customer attributes using unsupervised machine learning techniques. In addition, one complementary model is presented which indicates the effect of the nature of customer information on the quality of cluster and its interpretability. This study particularly laid focus on the customer behavioural attributes, which are nominal (categorical) in type. Boriah et al. (2014) asked the question "Which similarity measure is best suited for my data mining task?", Their experimental outcome suggested that there is no sole best performing similarity measure. Dataset used in this research is symmetric binary variables whose outcome can obtain only two values, i.e. 0 and 1. Binary variables often called symmetric, if there is no specific choice for the outcome, i.e. both outcomes are equally valuable and assigned identical weight when proximity measure is calculated. After a critical evaluation of the literature, the segmentation of symmetric binary dataset is executed in two-stages first via hierarchical clustering to determine the number of clusters followed by K-modes.

Dutt et al. (2017) stated that recently educational institutes initiate to aggregate and store the voluminous amount of data, such as the interaction of student on the university website, attendance record and student enrolment. Due to swift growth in educational data points, it needs a sophisticated set of algorithms to render insights. These complexities lead to

the development of the EDM field (Educational data mining). The dataset used in this study has a similar characteristics to that of educational data mining. Further, Dutt et al. (2017) stated numerous studies on educational data mining focused on the application of numerous data mining technique to the educational attributes. In their research, they did systematic literature review over a period of three decades, i.e. from 1983 to 2016 to determine the usability and applicability of data mining techniques in the domain of EDM. There are in total 35 such studies. On careful observation, the it was found that none of the studies used hierarchical clustering in combination with K-modes as a clustering algorithm because these studies used either metric/numerical datasets or mixed (Categorical or mixed) datasets. From that it can be inferred, this specific combination of hierarchical clustering followed by K-modes particularly for the symmetric binary dataset with low dimensionality and low volume is implemented for the first time in the field of educational data mining. Hence, it's a significant contribution of this study.

5.2.2. Practical Implication

The result of this paper indicates the presence of two categories of customer behavioural attribute. They are macro-behaviours and micro-behaviours. These categories have distinct context for behavioural attributes and for the rest of the parameters; behavioural source attributes, well-liked studies and preferred device type to navigate the website. For behavioural attributes, micro-behaviour denotes to non-dominant behavioural features whereas macro behaviour denotes dominant behavioural features. For the rest, macro-behaviour signifies the ordinal nature of behavioural features, i.e. it does not give significance to the magnitude of differences between features, but micro-behaviour takes into consideration the metric nature of features. However, in both cases, macro-behaviours are partially overlapping or similar between the cohorts/profiles. Therefore, formulating marketing communication strategies for campaigning based upon this can negatively influence the effectiveness of behavioural targeting. However, micro-behaviour varies across each profile, and they have a unique pattern, which helps the marketer to formulate strategies, for specific profiles, which enhances the relevance of advertisements and thus serve the purpose of efficient behavioural targeting using complex data and unconventional approach. In addition, these micro-behaviours succour to modify the current remarketing campaign which can prove pivotal in increasing the effectiveness and efficiency of the targeted advertisements.

Another objective of this study is to empower the SMEs to efficiently execute the behavioural targeting. As stated in Chapter 1, major limitations for SMEs are their tight budget constraints which don't allow them to purchase behavioural targeting services and even if they manage to buy the services, there is always a threat to the integrity of data (third-party). Further, surging variety and volume of the data along with complex machine learning algorithms raise their problems. Therefore, the introduced framework is for unsupervised machine learning techniques for the dataset, which varies in terms of volume, ability to handle dimensionality and type (specifically categorical and numerical). Further, R language has been used in this study, which is an open-source software and it has more than 12,500 additional packages as of May 2018 ("R (programming language)", 2018) which allows users to easily execute the machine learning techniques. Hence in conjugation, the presented framework and R language are capable of creating meaningful profiles, even with a low financial budget.

5.2.3. Future Research and Research Limitations

This study laid a foundation for future research work in supervised machine learning especially classification. Classification is a supervised machine-learning algorithm. This system aims to generate a mapping (also called a hypothesis or model) between a given set of documents and class labels. It is then used to determine automatically the class of new unlabelled document. In this study, the profiles were identified along with the patterns exhibited by them pertaining to each profile, therefore prediction can be built for Indian website visitors of the UT who are interested in master studies, to classify the new visitors belongs to this group into one of the six discovered behavioural profiles. Furthermore, future research can be focused on specific customer behavioural attribute, which influences the conversion rate of profiles. For instance, behavioural profile 'Interested-HP' or 'Interested high potential prospects' of Indian website visitors interested in the master studies of the University of Twente have a strong influence of 'questions via web form' on the conversion (submission of application dossier). The 'question via web form' is often manifested when a visitor doesn't find the information regarding the specific queries on the university website. These questions are in the form of text and the volume of data generated every day is enormous. This immense quantity of data is often unstructured text; therefore, it cannot be easily processed and perceived by computers. Hence, it requires effective and efficient techniques to unravel the useful patterns. Text mining is a process of extracting the useful information from text, it can be obtained either from social networks and inside organizations or the Web, and it has gained substantial attention in the recent years (Allahyari et al.; 2017). By realizing such information, researchers can gain insight into behaviours of profiles, in turn, they can use it for marketing campaigns as well as for personalization of a website for a specific cohort of users.

The conclusion of this research is limited by volume and veracity of the dataset used. Veracity refers to the issue of validity, i.e. the accuracy of data for the intended use. Hence confirmatory research using the identical techniques on the website visitors of University of Twente interested in master studies can be conducted for a period of two-years, it could potentially resolve the volume as well as validity issue. Further, comparison of behavioural patterns of the 2-year dataset (2016-2018) with ten months of the dataset used in this study will reveal whether pattern remains constant or alters with time. If patterns alter, it will unveil the amount of variations in the patterns, thus aid the marketer to determine a suitable time frame to update data for its prediction model. However, the proposed framework can only be used for categorical or numerical dataset but not for mixed datasets. In literature, for instance, K-prototype is the pervasive technique for clustering of mixed datasets (Vijaya et al.; 2004). Also, the behavioural attribute available for dataset determines the accuracy of the conclusion. However, the variety of raw data sources limits this research.

6. Reference

Adomavicius, D., & Tuzhilin, A. (2006). Personalization technologies: A process-oriented perspective. *Business Informatics* , *48* (6), 449-450

Ahmed, A., Low, Y., Aly, M., Josifovski, V., & Smola, A. J. (2011, August). Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 114-122). ACM.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Appiah, K., Hunter, A., Dickinson, P., & Meng, H. (2012). Implementation and applications of tri-state self-organizing maps on FPGA. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(8), 1150-1160.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243-256.

Arshad, A., & Ameen, K. (2015). Usage patterns of Punjab University Library website: a transactional log analysis study. *The Electronic Library*, *33*(1), 65-74.

Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(4), 340-350.

Augustijn, E. W., & Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International journal of health geographics*, *12*(1), 60.

Baranowska, M. (2014). Marketing theory. Behavioural segmentation. Retrieved from https://www.slideshare.net/monikaba5/marketing-theory-behavioural-segmentation

Bennett, S. C. (2010). Regulating online behavioral advertising. *J. Marshall L. Rev.*, *44*, 899.

Blackboard. (2014). Four Leading Strategies To Identify, Attract, Engage, and Enroll the Right Students, 1–7.

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2017). Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*, *46*(3), 363-376.

Boratto, L., Carta, S., Fenu, G., & Saia, R. (2016). Using neural word embeddings to model 77 user behavior and detect user segments. *Knowledge-Based Systems*, *108*, 5–14.

Boratto, L., Carta, S., Fenu, G., & Saia, R. (2016). Using neural word embeddings to model user behavior and detect user segments. *Knowledge-Based Systems*, *108*, 5-14.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (1999, September). Optics-of: Identifying local outliers. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 262-270). Springer, Berlin, Heidelberg.

Cao, L. (2010). In-depth behavior understanding and use: The behavior informatics approach. *Information Sciences*, *180*(17), 3067–3085. https://doi.org/10.1016/j.ins.2010.03.025

Cao, L. (2014). Behavior informatics: A new perspective. *IEEE Intelligent Systems*, *29*(4), 62–80. https://doi.org/10.1109/MIS.2014.60

Cao, Y., & Li, Y. (2007). An intelligent fuzzy-based recommendation system for consumer electronic products. *Expert Systems with Applications*, *33*(1), 230-240.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert systems with applications*, *34*(4), 2754-2762.

Chen, K., & Liu, L. (2009). "Best K": critical clustering structures in categorical datasets. *Knowledge and information systems*, *20*(1), 1-33.

Chester, J. (2012). Cookie Wars: How New Data Profiling andTargeting Techniques Threaten Citizens and Consumers in the"Big Data" Era. *European Data Protection: In Good Health?* 53-77. doi:10.1007/978-94-007-2903-2_4

Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied psychological measurement*, *11*(4), 329-354.

De Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, *324*, 126–145.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991-16005.

Etminani, K., Delui, A. R., Yanehsari, N. R., & Rouhani, M. (2009, July). Web usage mining: Discovery of the users' navigational patterns using SOM. In *Networked Digital Technologies, 2009. NDT'09. First International Conference on*(pp. 224-249). IEEE.

(Ed.). (2017, September). The economic value of behavioural targeting in digital advertising. Retrieved from ttps://datadrivenadvertising.eu/wp-content/uploads/2017/09/BehaviouralTargeting_FINAL.PDF

FACTS & FIGURES. (2018). Retrieved July 16, 2018, from https://www.utwente.nl/en/facts-and-figures/education/#key-figures

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing* Pandove, *2*(3), 267-279.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Formann, A. K. (1984). *Die latent-class-analyse: Einführung in Theorie und Anwendung*. Beltz.

Frolov, A. A., Husek, D., & Polyakov, P. Y. (2014). Two expectation-maximization algorithms for Boolean factor analysis. *Neurocomputing*, *130*, 83-97.

Goldfarb, A., & Tucker, C. E. (2011). Privacy Regulation and Online Advertising. *Management Science*, *57*(1), 57–71.

Gong, X., Guo, X., Zhang, R., He, X., & Zhou, A. (2013, December). Search behavior based latent semantic user segmentation for advertising targeting. In *2013 IEEE 13th International Conference on Data Mining* (pp. 211-220). IEEE.

Greenacre, M. (2010). Correspondence analysis of raw data. *Ecology*, *91*(4), 958-963.

Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In *ACM Sigmod Record* (Vol. 27, No. 2, pp. 73-84). ACM.

Guha, S., Rastogi, R., & Shim, K. (1999, March). ROCK: A robust clustering algorithm for categorical attributes. In *Data Engineering, 1999. Proceedings., 15th International Conference on* (pp. 512-521). IEEE.

Guideline privacy rules: protection or personal data in scientific research. (nd). Retrieved July 15, 2018, from https://www.utwente.nl/en/cyber-safety/cybersafety/privacy/guideline-for-research/

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, *17*(2-3), 107-145.

Ham, C. D., & Nelson, M. R. (2016). The role of persuasion knowledge, assessment of benefit and harm, and third-person perception in coping with online behavioral advertising. *Computers in Human Behavior*, *62*, 689-702.

Hosseni, M, .B, Tarokh M. J. (2011). Customer Segmentation Using CLV Elements. *Journal of Service Science and Management* 04(03): 284–290.

Hsu, C. C. (2006). Generalizing self-organizing map for categorical data. *IEEE transactions on Neural Networks*, *17*(2), 294-304.

Huang, J. J., Tzeng, G. H., & Ong, C. S. (2007). Marketing segmentation using support vector clustering. *Expert systems with applications*, *32*(2), 313-317.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *DMKD*, *3*(8), 34-39.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *DMKD*, *3*(8), 34-39.

J. Gower. Similarity, dissimilarity and distance, measures of. Encylcopedia of statistical sciences, 2004.

Jain, B. J. (2016). Homogeneity of Cluster Ensembles, 1–29.

Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Trans Knowl Data Eng* 2004, 16:1370–1386.

Jolliffe, I. T. (2002). Graphical representation of data using principal components. *Principal component analysis*, 78-110.

Kailing, K., Kriegel, H. P., Kroeger, P., & Wanka, S. (2003, September). Ranking interesting subspaces for clustering high dimensional data. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 241-252). Springer, Berlin, Heidelberg.

Karpatne, A., Khandelwal, A., Boriah, S., & Kumar, V. (2014, April). Predictive learning in the presence of heterogeneity and limited training data. In *Proceedings of the 2014 SIAM International Conference on Data Mining* (pp. 253-261). Society for Industrial and Applied Mathematics.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*(8), 68-75.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1-3), 1-6.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Cluster analysis in industrial market segmentation through artificial neural network. *Computers & Industrial Engineering*, *42*(2-4), 391-399.

Lambrecht, A., & Tucker, C. (2013). When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, *50*(5), 561-576.

Lee, S. C., Suh, Y. H., Kim, J. K., & Lee, K. J. (2004). A cross-national market segmentation of online game industry using SOM. *Expert systems with applications*, *27*(4), 559-570.

Liu, D., & Graham, J. (2018). Simple Measures of Individual Cluster-Membership Certainty for Hard Partitional Clustering. *The American Statistician*, (just-accepted), 1-25.

Lourenco, F., Lobo, V., & Bacao, F. (2004). Binary-based similarity measures for categorical data and their application in Self-Organizing Maps.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Meredith, S. (2018, April 10). *Facebook-Cambridge Analytica: A timeline of the data hijacking scandal*. Retrieved from CNBC: https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

Mishra, B. K., Rath, A., Nayak, N. R., & Swain, S. (2012, August). Far efficient K-means clustering algorithm. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 106-110). ACM.

Oja, M., Kaski, S., & Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. *Neural computing surveys*, *3*(1), 1-156.

Palla, K., Ghahramani, Z., & Knowles, D. A. (2012). A nonparametric variable clustering model. In *Advances in Neural Information Processing Systems* (pp. 2987-2995).

Pandey, S., Aly, M., Bagherjeiran, A., Hatch, A., Ciccolo, P., Ratnaparkhi, A., & Zinkevich, M. (2011). Learning to target. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11*, 1805. https://doi.org/10.1145/2063576.2063837

Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *12*(2), 16.Milligan, G. W., &

Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *12*(2), 16.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications*, *36*(2), 3336-3341.

Punj, Girish, and David W. Stewart. "Cluster analysis in marketing research: Review and suggestions for application." *Journal of marketing research* (1983): 134-148.

R (programming language). (2018, June). Retrieved July 16, 2018, from https://en.wikipedia.org/wiki/R_(programming_language)

Saia, R., Boratto, L., Carta, S., & Fenu, G. (2016). Binary sieves: toward a semantic approach to user segmentation for behavioral targeting. *Future Generation Computer Systems*, *64*, 186-197.

Santana, A., Morais, A., & Quiles, M. G. (2017, May). An alternative approach for binary and categorical self-organizing maps. In *Neural Networks (IJCNN), 2017 International Joint Conference on* (pp. 2604-2610). IEEE.

Singal, H., Kohli, S., & Sharma, A. K. (2014). Web analytics: State-of-art & literature assessment. *2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)*, 24–29.

Smit, E. G., Van Noort, G., & Voorveld, H. A. (2014). Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe. *Computers in Human Behavior*, *32*, 15-22.

Srimani, P. K., & Srinivas, A. (2011, December). Behavioral Targeting—Consumer Tracking. In *AIP conference proceedings* (Vol. 1414, No. 1, pp. 56-60). AIP.

Strong, E. K. (1925). *The psychology of selling and advertising*. McGraw-Hill book Company, Incorporated.

Stroud, Dick. (2006). Customer Intelligence. *Journal of Direct Data and Digital Marketing Practice* 7(3): 286–288.

Tamasauskas, D., Sakalauskas, V., & Kriksciuniene, D. (2012, December). Evaluation framework of hierarchical clustering methods for binary data. In *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on* (pp. 421-426). IEEE.

Tsiptsis, K., & Chorianopoulos, A. (2009). *Data Mining Techniques in CRM: Inside Customer Segmentation*. Wiley.

Tu, S., & Lu, C. (2010, November). Topic-based user segmentation for online advertising with latent dirichlet allocation. In *International Conference on Advanced Data Mining and Applications* (pp. 259-269). Springer, Berlin, Heidelberg.

Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, *51*(5), 546-562.

Usage statistics and market share of Google Analytics for websites. (2018, July). Retrieved July 16, 2018, from https://w3techs.com/technologies/details/ta-googleanalytics/all/all

Vijaya, P. A., Murty, M. N., & Subramanian, D. K. (2004). Leaders–Subleaders: An efficient hierarchical clustering algorithm for large data sets. *Pattern Recognition Letters*, *25*(4), 505-513.

Wedel, M., & Kamakura, W. A. (2000), "Market segmentation: Conceptual and methodological foundations" (2nd ed.). *Dordrecht: Kluwer*.

Wijaya, B. S. (2015). The development of hierarchy of effects model in advertising. *International Research Journal of Business Studies*, *5*(1)

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). Citeseer.

Wu, X., Yan, J., Liu, N., Yan, S., Chen, Y., & Chen, Z. (2009). Probabilistic Latent Semantic User Segmentation for Behavioral Targeted Advertising*. *Third International Workshop on Data Mining and Audience Intelligence for Advertising*, 10–17.

Wu, X., Yan, J., Liu, N., Yan, S., Chen, Y., & Chen, Z. (2009, June). Probabilistic latent semantic user segmentation for behavioral targeted advertising. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising* (pp. 10-17). ACM.

Yao, Z., Holmbom, A. H., Eklund, T., & Back, B. (2010, July). Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis. In *Industrial Conference on Data Mining* (pp. 292-307). Springer, Berlin, Heidelberg.

Zhao and L. Xue. 2013. Competitive target advertising and consumer data sharing. *Journal of Management Information Systems* 29, 3, 189–222.

Zhao. 2012. Service design of a customer data intermediary for competitive target promotions. *Decision Support Systems* 54, 1, 699–718.

Zuiderveen Borgesius, F. (2015). Improving privacy protection in the area of behavioural targeting.

7. Appendixes

Appendix 1

Figure 19: Illustrates the outcomes of the performance of the (ten) hierarchical clustering algorithm on symmetric distance measurement

TABLE 7. RESULTS OF SYMETRIC DISTANCE MEASUREMENTS

Error	hamming	dmatch	dsqmatch	rt	ss1
Average	3.3%	3.3%	3.3%	3.3%	2.5%
Centroid	3.3%	3.3%	3.3%	3.3%	3.3%
Complete	**1.7%**	**1.7%**	**1.7%**	**1.7%**	**1.7%**
Density	49.2%	49.2%	49.2%	49.2%	49.2%
Flexible	5.0%	**1.7%**	5.0%	2.5%	5.0%
Mcquitty	45.8%	45.8%	45.8%	45.8%	45.8%
Median	49.2%	49.2%	49.2%	49.2%	41.7%
Single	49.2%	49.2%	49.2%	49.2%	49.2%
Twostage	2.5%	4.2%	2.5%	2.5%	2.5%
Ward	3.3%	**1.7%**	3.3%	2.5%	2.5%

In Figure 19, accuracy of each algorithm is analysed by approximating the ratio between the counts of errors where algorithm/ method assigns wrong cluster and total count of investigated cluster. It implies lower the number better is the accuracy of the algorithm.

Appendix 2

Figure 20: Summary of nature of information of user profile types by Cufoglu (2014)

Table 1. Comparison of the User Profile Types

User Profile Type	Description	Techniques Used	Advantages	Disadvantages
Explicit User Profiles	User manually creates user profile	Questionnaires, Rating	Information gathered is usually of high quality	Requires a lot of efforts from user to update the profile information
Implicit User Profiles	System generates user profile from usage history of interactions between user and content	Machine learning algorithms	Minimal user effort is required and easily updatable by automatic methods	Initially requires a large amount of interaction between user and content before an accurate user profile is created
Hybrid User Profiles	Combination of explicit and implicit user profiles	Both explicit and implicit techniques	To reduce weak points and promote strong points of each of the techniques used	N/A

Appendix 3

Figure 21: Synopsis of user profile methods by Cufoglu (2014)

Table 2. Comparison of User Profiling Methods

User Profiling Method	Description	Techniques Used	Advantages	Disadvantages
Content-based Filtering	Filtering content from a data stream based on extracting content features that have been expressed in	Vector Space model, Latent semantic indexing, Learning information agents, Neural network agents	Objective analysis of large and/or complicated (e.g. multimedia) sources of digital material without much user involvement	1. Content dependent 2. Hard to introduce serendipitous recommendations as approach suffers from tunnel vision effect
Collaborative Filtering	Filtering items based on similarities between target users collaborative profile and peer user/group	Memory-based and Model-based	1. Content independent 2. Proves more accurate than content-based filtering for most domains of use enables introduction of serendipitous choices	1. Sparsity: poor prediction capabilities when new item is introduced to database due to lack of ratings 2. First-rater: poor recommendations made to new users until they have enough ratings in their profiles for accurate comparison to other users
Hybrid Filtering	Combines two filtering techniques	Collaborative Content based	To reduce weak points and promote strong points of each of the techniques used	Weak points can out-weight strong points if the hybrid is created naively

Appendix 4

Figure 22: Classification of clustering algorithms with respect to big data characteristics by Fahad et al. (2014)

Categories	Abb. name	Volume			Variety	
		Size of Dataset	Handling High Dimensionality	Handling Noisy Data	Type of Dataset	Clusters Shape
Partitional algorithms	K-Means [25]	Large	No	No	Numerical	Non-convex
	K-modes [19]	Large	Yes	No	Categorical	Non-convex
	K-medoids [33]	Small	Yes	Yes	Categorical	Non-convex
	PAM [31]	Small	No	No	Numerical	Non-convex
	CLARA [23]	Large	No	No	Numerical	Non-convex
	CLARANS [32]	Large	No	No	Numerical	Non-convex
	FCM [6]	Large	No	No	Numerical	Non-convex
Hierarchical algorithms	BIRCH [40]	Large	No	No	Numerical	Non-convex
	CURE [14]	Large	Yes	Yes	Numerical	Arbitrary
	ROCK [15]	Large	No	No	Categorical and Numerical	Arbitrary
	Chameleon [22]	Large	Yes	No	All type of data	Arbitrary
	ECHIDNA [26]	Large	No	No	Multivariate Data	Non-convex
Density-based algorithms	DBSCAN [9]	Large	No	No	Numerical	Arbitrary
	OPTICS [5]	Large	No	Yes	Numerical	Arbitrary
	DBCLASD [39]	Large	No	Yes	Numerical	Arbitrary
	DENCLUE [17]	Large	Yes	Yes	Numerical	Arbitrary
Grid- based algorithms	Wave-Cluster [34]	Large	No	Yes	Special data	Arbitrary
	STING [37]	Large	No	Yes	Special data	Arbitrary
	CLIQUE [21]	Large	Yes	No	Numerical	Arbitrary
	OptiGrid [18]	Large	Yes	Yes	Special data	Arbitrary
Model- based algorithms	EM [8]	Large	Yes	No	Special data	Non-convex
	COBWEB [12]	Small	No	No	Numerical	Non-convex
	CLASSIT [13]	Small	No	No	Numerical	Non-convex
	SOMs [24]	Small	Yes	No	Multivariate Data	Non-convex