UNIVERSITY OF TWENTE.

Faculty of Behavioral, Management, and Social Science

Discovering behavioral profiles for website visitors of higher educations

Alireza Sadeghi

Thesis Assignment
M.Sc. Business Administration - Strategic Marketing and Business Information
April 3rd 2018

Supervisors:
Dr. S.A De Vries
Dr. R.P.A. Loohuis

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

Abstract

Several studies made attempts at using behavioural data to support business activities. Some studies attempted to creating user profiles to identify potential customers to increase customer spending and others have attempted to find the best performing algorithm to increase the accuracy of preceding studies. The goal of this paper is to develop a framework for customer profiling and customer attribute prediction within the marketing context using Machine Learning and Customer attributes as well as developing a *multi-step user profiling process* model for user profiling. The goal in this paper is to discover behavioural profiles of website visitors for higher educations. Previous studies used Machine Learning and customer attributes to identify the most profitable profiles of existing customers for the purpose of increasing spending amount but this paper focuses on identifying behavioural profiles to increase customer base and increase conversion rate. Thus, this study focusses on finding behavioural profiles within website visitors for higher education by utilizing behavioural data and applying proposed model and framework in this paper. The outcome provides insight for University of Twente marketing department as to what behaviours lead to higher conversion. The paper proposes a framework and a model, where the framework provides a guideline for different research goals based on customer attributes & Machine Learning algorithms and the model provides a guideline on the way customer data should be processed to gain a profound insight from data. The analysis reveals three behavioural profiles for the website visitors of the University of Twente by utilizing the framework and the model proposed in this paper. The outcome provides evidence that the outcome is more profound when the proposed framework in combination with proposed model is used compared to the previously one-step user profiling used in the literature.

Keywords: behavioural targeting; machine learning; behavioural profiling

Table of Contents

Definition of terms

Behavioural targeting: Behavioural targeting or BT is a technique used by marketers, which belongs to the branch of targeted advertisement. In BT, marketers make use of users data such as visited websites, the amount of time spent on each website or otherwise known as browsing behaviour to target the most appropriate visitors/customers (Chen & Stallaert, 2014). Behavioural targeting is also known as Online behavioural advertising (OBA).

Cookie: is a small file that is downloaded into computers when users access certain websites. These files enable websites to identify user's computers. Cookies carry a modest amount of data, which can be accessed either by web server or the client computer (FESBAL, 2013). They have a number of applications, which one of them is in Behavioural targeting. By using the user-specific information, content and advertisement are tailored to interests of individuals visitor(Bureau, 2014).

Conversion rate: Conversion rate is the process of turning a website visitor into a paying customer. The usage of this term is contingent on the nature of websites where some consider it to be a result other than sales("Conversion Rate," 2017).An example of desired actions includes but is not limited to memberships, registration, newsletter subscriptions. A way of increasing the interest level of a visitor is to match with the right visitor or another way around, adjust the website content to the visitor ("Conversion Rate," 2017).

Click-through rate (CTR): is the rate where the paid-per-click advertisement is clicked. Often it is a measurement way of advertisement success. Normally, a high CTR indicates that the advertisement is been relevant to the chosen audience("Click-through rate (CTR): Definition," 2017). A result, it is often used in measuring the success of BT advertisement. It is computed from the number of times that an advertisement has been clicked divided by the number of times an advertisement has been viewed (Kim, 2017).

CLUTO: is a software package that enables clustering of low and high dimensional data set and the characteristic of various clusters(“CLUTO - Software for Clustering High-Dimensional Datasets,” 2006).

1. Introduction

Nowadays, organizations have access to vast amount of data about their customers and utilizing such data in a meaningful way, could excel marketing processes (Terradata, 2015). However, the majority of marketers consider online data as one of the most underutilized sources of information in organizations (Terradata, 2015). In addition, the number of people who use the Internet in the world has grown by almost 69% from 2010 to 2016 (Internet Users, 2017). If this growth is by any means an indication of the user-generated data velocity, then the importance of utilizing such data must be emphasized and explored.

Moreover, organizations are overwhelmed by the volume and velocity of generated customer data or in general Big Data. They are often unable to gain any meaningful insight from Big Data. For this reason, it is important to outline an overview of approaches to knowledge discovery from Big Data. Such overview assists organizations to gain valuable insight from various Big Data sources so that it could be used to support decision making in various business areas, especially in the marketing field. It is well known that much of the marketing effort and budget is wasted on the wrong audiences. Therefore, by applying appropriate approaches to customer data, marketing departments can readjust their effort to be more efficient and effective. Furthermore, many SME's are faced with financial or resource limitations (such as time or personnel limitation) for marketing and advertising (Blackboard, 2014). Consequently, organizations often use BT tools in a very generic way or as an alternative to offline marketing.

The premise of BT is to distinguish the individualistic differences (whether behaviour or interest) between two seemingly identical customers. Such differences can only be identified when customers are analyzed in a more complex manner that yields profound yet understandable insight. Traditional segmentation methods are too crude and ignore such individualistic differences. Currently, user profiling and segmentation are mainly done by

means of *target audience persona creation* based on marketing tools like surveys. Such approaches to user profiling do not acknowledge the difference in seemingly identical customers, which is often the result of focusing and relying on explicit customer information. However, Behavioral targeting incorporates implicit customer data, which allows marketers to realize subtle individualist differences and leverage them to deliver more relevant advertisements to relevant potential customers. Often individualistic differences are discovered as a result of analysing implicit user data such as user browsing history and search behaviour (Chen & Stallaert, 2014).

Therefore, defining a framework of BT can empower SMEs and as a result, it could potentially lower their entry barrier due to resource limitations. Moreover, a good understanding of various approaches on how to leverage customer data can assist all businesses in decision makings. In addition, insight generated by analysing Big Data using proposed models in this paper could lead to the discovery of underserved customer segment(s), where an appropriate analysis such as sequence analysis can lead to the development of a plan that attends the needs of underserved customers.

A great deal of existing research is about the effectiveness, efficiency, accuracy of different behavioural techniques and its value in terms of improving online advertising. For instance, Goldfarb & Tucker (2011) observed that users are less likely to buy products, after viewing an advertisement that is not interest-based targeted. Furthermore, (Yan et al., 2009)state BT increases the effectiveness of advertisement with the measure of CTR, by 670%. Other researchers of BT addressed different segmentation attributes, techniques, methods and the ways algorithms improve efficiency and accuracy of segmentation. However, none have outlined an overview of various approaches to user profiling (also known as segmentation) and prediction based on Machine Learning and customer attributes. In addition, prior research has

failed to acknowledge the different types of customer data and the processing sequence of such data for user profiling.

The goal of this paper is to develop a framework for customer profiling and customer attribute prediction, within the marketing context using Machine Learning and customer attributes, as well as developing a *multi-step User profiling process* model for user profiling. In addition, the aim is to discover behavioural profiles of website visitors for higher educations based on data-driven characteristics. The outcome would ideally reveal the behavioural profiles of website visitors for the University of Twente. Such profiles provide insight about (offline and online) behavioural patterns of potential visitors. In turn, future research could use such insight to identify the granular patterns and behavioural sequences of desired groups, to create customised marketing campaigns.

Research Question: What are the behavioural profiles of website visitors in higher education?

In order to answer the main research question, the following questions need to be addressed:

I. What customer attributes can be used for profiling?
II. How can behavioural profiles be identified?
III. Are discovered behavioural user profile consistent when controlling for different factors?

In order to realise this paper's objective, the relevant literature regarding segmentation such as user profiling and, Machine Learning algorithms are reviewed. However, the core literature for this study is user segmentation, user profiling and finding appropriate attributes and methods for behavioural user profiling to enhance online targeting strategies.

The data used in this paper are from the University of Twente, therefore making this paper a case study. Therefore, this paper provides insight into how visitors' data of the University of Twente can be leveraged to create behavioural profiles. This paper is an explorative case study, it focuses on behaviours of website visitors of the University of Twente. The data used in this

paper is secondary data of the University of Twente' website and its business data (CRM database).

This paper attempts to close the gap in the literature by developing a framework on various approaches on how customer attributes can be leveraged to gain insight for supporting business activities, specifically for user profiling and prediction for use in online targeting approach. The application for such analysis in marketing field is to gain profound and meaningful insight on visitors' groups with similar characteristic and use such insight to create accurate and efficient BT campaigns. Furthermore, this study lay the foundation for future research, whereby analysis on a desirable profile could provide a detailed understanding of visitors in terms of the sequence and temporal behaviour manifested.

The paper is organised into 5 chapters and it is structured as follows. The next chapter covers the theoretical framework, which is the literature review of prior research on topics such as Behavioural targeting, Machine learning algorithm, knowledge discovery in Big Data and user segmentation. Next chapter outlines the methodology of this paper. It expands on the nature of data and its collection method, analysis strategy and (behavioural) attributes used to conduct the analysis. The following chapter describes the result of all analyses, where the outcome of each analysis is presented. Within the same chapter, the results are visualized in a side by side manner to assist interpretation. In the last chapter, conclusions and discussions are presented as well as the limitation and theoretical and practical implication of this paper.

2. Theoretical framework

In order to achieve the goals of this paper, various aspects need to be outlined such as the definition of behaviour, various Machine Learning algorithm, customer attributes. This chapter presents various core literature used in this paper. Each section represents a relevant aspect of core literature that allows the researcher to propose the framework and model presented at the end of the chapter and to conduct its analysis and to propose. As result, the relevant literature regarding each aspect is summarized and discussed briefly in each section.

2.1. Definition of Behaviour

Before being able to discover the Behavioural user profiles, one needs to understand and define behaviour in the first place. Therefore, here in this section, the definition of behaviour for this paper is elaborated. In the traditional sense, a behaviour is the manner whereby a system or a being interact or react to another one. It is known by the actions and manner that such beings interact with their environment (Cao, 2014). Behaviours in the non-digital world have been vastly studied from different aspect due to their explicitness (Cao, 2014). However, with advancement in technology, behaviour takes very complex forms as it includes the digital implicit form such as the way individuals seek out information or react to the digital or physical environment. Behaviours which are recorded in digital form are often referred to as "Soft behaviour" (Cao, 2014). Cao (2014) refers to such behaviours in the digital era as "Behaviour Computing" or "Behaviour Information". He states that such behaviours consist of "Methodologies, techniques and practical tools for representing, modelling, analysing, learning, discovering, and utilizing human, or animal, organization, social, artificial , and virtual behaviours, behavioural interaction and relationships, behavioural networks , behavioural patterns and behavioural impacts" (Cao & Yu, 2012).

In the field of behaviour 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". Understudying behaviour computing provide opportunity improvement and discovery of certain behaviours/behavioural pattern that could be used for different application for management and business intelligence (Cao, 2014). Cao & Yu (2012) propose a different way of looking at the behaviours of individuals. Many researchers used to look and study behaviours in a qualitative manner, however, Cao & Yu (2012) propose a quantitative way of studying behaviour in order to discover knowledge.

Fayyad, Piatetsky-Shapiro, & Smyth (1996) describe the pattern as "an expression in some language describing a subset of the data or a model applicable to the subset". They underline that the unravelled patterns must be valid to some degree on new data, be novel and understandable that could provide useful information that benefits users and/or tasks. Therefore, they conclude that for any pattern to be considered as knowledge, it needs to pass a certain threshold to provide useful information (Fayyad et al., 1996).

In conclusion, this paper uses the definition of behaviour proposed 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 example of behaviour includes actions and operations that visitors manifest while browsing the website of higher education in order to gather information. A bundle of behaviours (either in the digital or physical environment) represent a behavioural pattern of website visitors, which is how behavioural user profiles are depicted in this paper. To discover behavioural profiles (which is based on behavioural patterns) of website visitors for higher educations, a certain technique needs to be applied to behavioural data in order to extract meaningful insight. The next section summarizes the fundamentals of such techniques (generally is known as *Knowledge Discovery* processes) that allows information to be extracted from raw databases, which in this study is the behavioural data of website visitors.

2.2. Knowledge Discovery

As mentioned in the previous section, certain techniques need to be applied to data (which in this study is visitor's behavioural data) to extract meaningful yet profound informative insight. Fayyad et al. (1996) coined the term *Knowledge Discovery* that compromises the collection of aforementioned techniques and segregated it into two main categories. This section elaborates on fundamental categories of Knowledge Discovery, which assists in identifying appropriate techniques for the research goal, namely segmentation and prediction of customer attributes.

Fayyad et al. (1996) define two main broad categories of knowledge discovery namely, *Verification* and *Discovery*. The first category, namely Verification, is used to prove or disprove a hypothesis. The second category, namely Discovery is used to discover pattern within data. Moreover, Fayyad et al. (1996) segregate the Discovery category into two sub-categories of *Prediction* and *Descriptive*. In the Prediction sub-category, patterns and variables (such as behaviour or spending amount) are used to predict a future event whereas the Descriptive sub-category unravel the naturally occurring patterns in a way that is untestable for an analyst with use of various methods and techniques (Fayyad et al., 1996).

In conclusion, two sub-categories relevant for this study are Predictive & Descriptive. The techniques in Descriptive sub-category can be used for segmentation, where techniques in the Predictive sub-category can be used for customer attribute prediction. The main techniques for the Predictive and Descriptive sub-categories are outlined in the following section.

2.3. Knowledge Discovery techniques

This section describes the main (Machine Learning algorithm) techniques used for the two subcategories of Discovery, namely Descriptive and Predictive. Outlining (Machine Learning) techniques of each aforementioned category expands on the previous section by describing the technical aspects and application of each technique. Choosing one over another technique is

completely dependent on the goal of each individual research. However, in order to provide an overview of different approaches later in this chapter, techniques from both categories are elaborated here.

Generally, Machine Learning techniques are divided into two main categories of *Unsupervised* and *Supervised* (Doig, 2015). The Unsupervised Machine Learning corresponds to the Descriptive sub-category described by Fayyad et al.(1996). In this category of Machine learning the algorithm finds natural occurring patterns among data. The second category, namely Unsupervised, corresponds to Predictive subcategory described by Fayyad et al. (1996). This category of Machine Learning algorithm classifies or identifies certain events or groups of people within the database based on certain features. This type of Machine Learning uses the occurrence of certain (desirable) events , based past historical data, to predict a future event. Here below, first the Unsupervised Machine Learning techniques are stated followed by Supervised Machine Learning techniques.

2.3.1. Unsupervised Machine Learning techniques

Clustering techniques and its variation are Unsupervised Machine Learning techniques, which are often used to find the natural or arbitrary structural pattern in data is determined using distances between data entries. The *Non-probability* variation of this technique allows cases to belong to only one group at a time and it is often referred as *Hard clustering*. Depending on the objective and the goal of an analysis, this could be considered as a limitation of Non-probabilistic variation of this technique. For instance, people have multiple dynamic interests where it changes over a period. However, Hard Clustering technique group visitors based on user behaviour at a certain collection time. This generates a rather a static image of visitor's behaviour that might be irrelevant after a while if the website content changes dramatically. Nevertheless, knowledge discovered as a result of processing data with Hard clustering technique provides valuable insights.

The two sub-types of Non-probabilistic variation of clustering technique are *Hierarchical* and *Non-hierarchical* (Hui, 2017). In the Hierarchical technique, prior knowledge about the number of groups is not required, as the algorithm can calculate the ideal number of clusters. However, this technique is not suitable for analysing large databases, as it requires high computational power. For analysing high-dimensional large datasets, Non-hierarchical such as *K-means* or *K-modes* is commonly used. However, such technique requires prior knowledge regarding the appropriate number of groups. Nonetheless, sometimes such knowledge on the number of groups is not known and therefore an additional test is required to determine the number of groups. An example of such test to determine the appropriate number of groups is '*within-cluster sum of squares*' also known as *Wcss* (Dao, Duong, & Vrain, 2015). This score measures the within-cluster distance of observations from their centroid. The aim of this test is to reduce the distance between observation in each cluster to a reasonable degree. The exact point number of groups is determined by visualization of the Wcss score on a Scree plot by employing the *Elbow method*. In this method, the appropriate number of groups is the point in the graph, where the Wcss score does not change dramatically (Asanka, 2017).

2.3.2. Supervised Machine Learning techniques

Classification is a (Supervised) Machine Learning algorithm that partition data set based on certain user pre-defined labels (Hand, 1981). Classification is a form of predictive technique that partition data into categorical variables. In order to predict a numerical, real-value variable a *Regression* technique (Multiple regression or Linear regression) is used to estimate the outcome based on new data. One of the popular Classification technique is called *Decision tree*. This technique makes use of recursive partitioning to divide the observations by data-driven threshold for each variable in multiple levels (Chorianopoulos, 2016). This technique could be used in allocating observation to various pre-determined segments.

This section outlines various Knowledge Discovery techniques described by Fayyad et al. (1996), which are known as Machine Learning techniques in the IT field. The two main types of Machine learning algorithms are Unsupervised and Supervised. Describing allows for better understanding of each technique, which also lays the foundation for identifying the appropriate technique to achieve the goal of this paper. However, further information is required for a good understanding of how segmentation and user profiling ought to be done. Therefore, the next section outlines the common approaches to segmentation based on the literature.

2.4. Segmentation approaches

In order to create user profiles, different approaches of segmentation must be realized. This is important as the outcome of user profiling should be simple and profound, yet not generic. Therefore, understanding various segmentation approaches allows for a realization of an approach that strikes the right balance, where outcomes are simple to understand and yet profound. In general, segmentation takes two main forms and they are as follow (Boratto, Carta, Fenu, & Saia, 2016; Dolnicar, 2008):

- *Apriori* (Common-sense)
- *Posteriori* (Post-hoc, Data-driven)

In the Posteriori approach, users are grouped based on data-driven similarities. So as, users are segmented into groups based on the between user's similarities. This approach generates a user profile outcome, that is not easily interpretable. However, it has the ability to reveal hidden relations among users that are overlooked by the typical (Common-sense) segmentation approach. A common technique for the Posteriori approach is Cluster analysis. In contrast, the Apriori approach does not take advantage of data-driven segmentation, thus runs the risk of superficial or generic clusters.

In an effort to address the shortcomings of the Apriori and Posteriori approaches, Dolnicar (2008) proposes a *Hybrid* segmentation approach. In this manner, segmentation is done in a

two-step approach by using the combination of two aforementioned approaches. Thus, his four-approach concept is as following:

- Apriori-Posteriori
- Posteriori- Apriori
- Apriori - Apriori
- Posteriori- Posteriori

These approaches are written in the sequence that they are ought to be applied and used. Meaning that in the first approach, the segmentation process beings with common-sense (Apriori) segmentation and then each segment is divided into more refined sub-segment by using Posteriori (Dolničar, 2004). Using the proposed approaches generates the benefit of potentially revealing hidden segments and provide easy interpretation to those segments.

In conclusion, there are three approaches to segmentation, namely *Apriori*, *Posteriori* and *Hybrid*. The first approach, Apriori, allows for a segmentation that is simplistic but logical and is often based on certain prior knowledge. The second approach, Posteriori, allows for a segmentation that can be sometimes counter-intuitive and profound, but difficult to understand. The third approach proposed by Dolnicar (2008), Hybrid, allows for a segmentation that utilizes the prior two approach to achieve a segmentation , which is simple to understand yet generates profound insight. The last approach is an interesting approach to segmentation. However, this approach fails to acknowledge how different user data can be used to generate a more realistic and accurate segmentation result. Therefore, the following section outlines various user profiling, based on different types of user data.

2.5. User profiling types

The previous section outlined and discussed various approaches to segmentation. However, as stated before, such approaches failed to account for the user data types in segmentation. This section expands on this by outlining user profiling types (also known as user segmentation types) based on different type of user data. Kanoje, Girase, & Mukhopadhyay (2014) describe two types of user profiling. One is based on the *Explicit* (traditional) user profiling, where websites ask users about their interest and preference to create or find their user profile. This is done to provide relevant content and setting to users. However, Explicit user profiling is often inaccurate since many users are unwilling to give out (personal) information due to privacy concerns.

The second type of user profiling, overcomes this issue by using *Implicit* data of users (Kanoje et al., 2014). This type relies on the interaction of users with the website or otherwise known in this paper as user behaviour. This type of data is richer source of user data, as it provides more detail about users. However, such user profiling using Implicit data might not always be easily interpretable, thus *Hybrid* user profiling is proposed to overcome interpretability issue of two aforementioned user profiling types.

Khosrow-pour (2009) proposes the third type, so-called *Hybrid* profiling, which is the combination of the previous two types. He states that the efficiency and accuracy of the user profiles are dependent on the quality and the amount of available data (Khosrow-Pour, 2009). By combining the two aforementioned types, user profiles would reflect more accurate and realistic preference of users. A summary of user profiling types can be found in appendix 1 (Khosrow-Pour, 2009).

In conclusion, the three user profiling types indicate that data types are important in creating meaningful and accurate user profiles. The Hybrid profiling addresses the shortcoming of user profiling by using Explicit and Implicit data, which is immensely important for creating an

accurate, simple and yet profound user profile. Nonetheless, realizing the user profiling approaches and its types is not sufficient to create accurate user profiles. To do so, user profiling methods should be realized and understood, which provides context to user profiles and assists in their interpretation. Thus, the next section outlines and describes various user profiling methods.

2.6. User profiling methods

Cufoglu (2014) proposes three user profiling methods. The first type is called *Content-based*, which assumes that a person behaves the same, under the same circumstance. Therefore, a user's behaviour is predicted from its past behaviour (Araniti, De Meo, Iera, & Ursino, 2003; GODOY & AMANDI, 2005; Kuflik & Shoval, 2000). The second type is called *Collaborative* (also known as collaborative filtering), which is based on assumption that users who exhibit similar behaviours belong to same groups (e.g. sex, age social class), or in other words belong to the same profiles (Araniti et al., 2003; GODOY & AMANDI, 2005; Kuflik & Shoval, 2000). This type basically states, that people with similar characteristics tend to be similar. Furthermore, Cufoglu (2014) proposes the third method, namely the *Hybrid* method, which takes advantage of combining the previous typed. This type generates accurate user profiles by revealing the true interests and preferences of users. Poo, Chng, & Goh (2003) segregated this type further into two main categories They distinguish two main categories for the *Hybrid* method, namely Static-profiling and Dynamic profiling. Poo, Chng, & Goh (2003) propose 4 sub-types for Hybrid method based on *Content-based* and *Collaborative*. They call recognize them as following: Static content profiling, Dynamic content profiling, Static collaborative profiling and Dynamic collaborative profiling. The four subtypes of the Hybrid method proposed by Poo et al. (2003) can be found in Appendix 2 and the summary of user profiling methods proposed by Cufoglu (2014) can be found in Appendix 3.

In conclusion, the *Hybrid* method is a comprehensive method that combines collaborative and content-based methods to utilize the strength of each user profiling method to overcome the shortcomings of each individual method. In addition, it accounts for the time perspective, by dividing *Hybrid* method into two sub-types of *Static* and *Dynamic.* Each sub-type allows for a more precise and accurate user profiling, that enables development of a user profiling strategy. Nonetheless, in order to create user profiles, various customer attributes need to be realized and outlined. Such attributes are outlined in the following section.

2.7. Customer attributes

In this section, various customer attributes used for user profiling in the literature are mentioned and defined briefly. It is important to know and recognize different customer attributes, as utilizing each or multiple of them yield a different image of users and as subsequent user profiles. Below, customer attributes used in the literature for user profiling can be found.

Demographic

This attribute enables advertisers to segment audience into meaningful target groups based on their demographic. Example of Demographic information are audiences' age, gender and income (Chen & Stallaert, 2014).

Geographic

Geographic or Geo-Targeting refers to a location attribute, which advertisement is based on the location of publisher's page or visitor in order to deliver the relevant advertisement. This is usually done based on the country of the visitor, city, postcode, IP address and other criteria (Plummer, Rappaport, Hall, & Barocci, 2007).

Psychographic

Psychographic attribute segregates potential users by their lifestyle, attitude, style and psychological traits. In turn, firms use such information to offer products tailored to the psyche of each user. Typical psychographic segmentation attributes are Interest, Opinion, Activities (IOA), activities and values. This approach is slightly different than that of Behavioural targeting attributes and it might be easily mistaken with BT. Psychographic approach divides users based on attributes such as personality traits, lifestyles, the degree of loyalty whereas BT attribute segregates customers based on information such as buying occasion, benefit sought (Directive Group, 2017).

Behavioural

Behavioural attributes allow for pattern recognition based on observed unconscious user-specific behaviour in a browsing session by analysing variables such as visit frequency, usage rate, benefit sought, occasion, user status, brand loyalty (Baranowska, 2014; Local Drirective, 2017). In turn, such attributes used to reveal patterns to group users with similar pattern together, otherwise stated, to identify narrower subgroups by combining other segmentation types (Local Drirective, 2017).

Daypart

This attribute, which was initially used in broadcasting, empowers marketers to break the day into several parts. Examples of Daypart segments are such as Morning drive, Daytime, Afternoon (Morning & Morning, 2017). By utilizing Daypart attributes, businesses are able to fine-tune their service to the desired audience. Each category of Daypart has an audience with specific characteristics, which is appropriate for certain products using different mediums (for Instance TV, Radio, Mobile). Therefore , businesses could refine their targeting by incorporating Daypart user attributes (Plummer et al., 2007).

In addition, advertising agencies are able to set a price for time slots and prioritize relevant content. As a result, knowledge could be discovered by analysing characteristics and the need of each Daypart segment (Leon, 2016). For instance, it is a known fact that the white-collar audience uses the internet a lot, it starts with a higher audience than television in the morning then soars up during mid-day and declines afterwards. Such information is very insightful in refining and targeting the desired audience (Plummer et al., 2007).

Affinity

Customers are known to have a preference and disposition towards certain brands, websites and services. In Affinity-targeting, the knowledge of customer's affinity is used to build indexes and profiles. In turn, such indexes and profiles are used for instance to find other users with similar affinity (Plummer et al., 2007). Furthermore, the knowledge of users' affinity can potentially unravel other characteristics that are associated with each brand or website. In the same manner, that viewers of certain news channels have common political views, it can be said that the political views or other desired views of individual users can be identified solely based on the type and frequency of visited websites.

Purchase-based

This attribute is used with the assumption that customers with similar purchasing characteristics tend to exhibit similar purchases (Dibb & Simkin, 1996). An example of Purchase-based attributes are items purchased and the monetary transaction associated with the history of customers. Furthermore, by utilizing such attributes businesses are able to create user profiles and use them to offer products or services that closely resemble users' characteristics (Tsai & Chiu, 2004). The utilization of Purchased-based attributes requires a business to understand its customer well, for building accurate Purchased-based profiles (Plummer et al., 2007).

All in all, there are 7 attributes mentioned in the literature, where each indicates a different aspect of customers. Realizing various customer attributes is necessary for developing an approach overview, which is given later in this chapter. In addition, it is interesting to realize various data sources that customer attributes could be retrieved from. The following section outlines various data sources.

2.8. Data sources

Understanding various customer attributes provides a solid foundation to draw a comprehensive overview of various approaches, for achieving the goals of this paper. Nevertheless, it is just as important to outline data sources of customer attributes. This knowledge assists researchers by providing a guideline on various data sources. In the literature 3 sources of data is mentioned, namely *Web data*, *Business data* and *Meta data*. Each data source is mentioned and briefly described below.

Web data

Web data typically comes in three forms. They are *Log files*, *Cookies* and *Query data* (Araya, Silva, & Weber, 2004). Cookies have many types and sub-types of its own and one of them is "First-party" Cookies. First-party Cookies are created by the same domain that a user visit. In this sub-type of Cookies, data is gathered by a single publisher (a single website) using different technologies such as *Cookies*, *tracking pixels* and an *agent*. Due to obvious reasons, the scope of information available from this sub-type of Cookies is limited. Data gathered by *First-party Cookies* often does not reflect the true and comprehensive interest or opinion of users. However, users' behaviour on a single website, provide its owner with knowledge and capability to improve for instance user experience (Plummer et al., 2007).

Query data is generated when visitors use the search function of websites. Each search term is gathered and stored in a file. Such data could be an indication of users' interest (Araya et al., 2004). The last sub-type of Web data is log files, which collects a comprehensive behavioural

data of users in detail (Araya et al., 2004). Although, this sub-type of Web data gathers detailed data on users, it is full of unnecessary data and contains quite a lot of noise. Therefore, usually, Cookies are a much easier and comprehensible source of data for an analysis.

Business Data

Back-end customer data can be used in combination with other sources of data to extract knowledge and understand behaviours associated with website navigation. Moreover, Businesses can identify and assign associate CRM profiles to profiles generated by Web data (Fennemore, 2011). According to a survey, about 55% of higher educations are not using CRM data for marketing and enrolment purposes (Blackboard, 2014). Business data include a range of information such as customer demographics and product information (Araya et al., 2004).

Meta Data

This type of data is generally used to describe data that embodies the structure and the content of websites (Araya et al., 2004). There are many Meta data features that can potentially provide knowledge by using it in combination with other data sources. Examples of sub-categories of Meta data include Structure data, Content data, Website network data and in general any data regarding the overview of the website and its pages (Araya et al., 2004). Furthermore, Content data mentioned earlier includes data such as images, brochures and free text within websites and their web pages (Araya et al., 2004).

In conclusion, each of the three aforementioned data sources provides different level of detail of user data. Understanding, various data sources and their scope provides knowledge that allows researchers to choose one or combination of data sources to achieve their research goal. However, in order to achieve the aim of this paper, the behavioural features used in the literature must be reviewed. Such knowledge provides insight as to what behavioural features have been previously used and so it provides knowledge as to the behavioural features are

appropriate for attaining the aim of this paper. The following section briefly discusses the behavioural features used by previous researchers.

2.9. Behavioural Features

As mentioned in the preceding section, knowing various behavioural features utilized by previous research, provide insight into the appropriate behavioural features for realizing the aim of this paper. Various research is dedicated to the study of the users' behavioural features for the purpose of user profiling (Boratto et al., 2016). Previous researchers made use of many features for the purpose of User profiling, by implementing various methods and techniques, which are described later. A number of users' behavioural features are mentioned below. As expressed before, the basic assumptions in using below-mentioned variables, are that users, who have similar features tend to manifest similar behaviours and vice versa (Dibb & Simkin, 1996; Tsai & Chiu, 2004).

Below, various behavioural features used in Behavioural targeting studies are mentioned (Baranowska, 2014; Castelluccia, 2012; Chopra, 2012; Deane, Meuer, & Teets, 2011; Dibb & Simkin, 1996; Jaworska & Sydow, 2008; Pandey et al., 2011; Plummer et al., 2007; Tsai & Chiu, 2004):

Website topic keyword
Types of visited website
Visited content
Frequency of visit
Time spent on each page
Time spent in each session
Searched keywords
Usage rate
Historical user activity
Purchasing activity
Navigational behaviour
Benefit sought
Brand loyalty
User status
Occasion
Time of logins
Date
Ads clicked
Clickstream
Device
Campaigns
Referring URL
Location
Sequence of page visited
Frequent mode of ad placement
In-text semantic

2.10. Framework for user profiling and customer attribute prediction

In this section, two models are developed to visualize strategies to user profiling based and customer attributes prediction, using two main categories of Machine Learning and customer attributes. The customer attributes are arranged from Explicit to Implicit order. The cells below each customer attribute (such as Demographics, Geographic) in figure 1, represent a strategy to either user profiling or customer attribute prediction. Each strategy is developed with a specific Machine Learning technique for the purpose of Knowledge Discovery from data. The first row represents Unsupervised Machine Learning algorithm such as clustering algorithm that allows User profiling. Clustering algorithms are generally divided into two main types, namely Hierarchical and Non-hierarchical. The basis for choosing such types depends on the volume and complexity of data, as shown in the model below. It is worth to point out that Non-hierarchical type has two sub-categories of Non-probability such as *K-means* and *K-modes* and Probability such as *Gaussian mixture* model and *Naïve Bayes*. However, the researcher of this paper chooses to focus on the Non-probability sub-category of Non-hierarchical.

			Explicitness <----------> Implicit						
Machine learning algorithm (ML)	Data type condition	ML Technique	Demographic	Geographic	Daypart	Affinity	Purchased-based	Behavioural	Psychographic
Unsupervised	**Low volume & low dimensional (<1000 entries)**	***Hierarchical Clustering + Non-hierarchical Clustering***	Segment users based on Demographics	Segment users based on their geographical location	Segmenting users based on time consumption/ Usage	Segmenting users based on similarity in brand affinity	Segmenting users based on their purchasing behaviour	Segmenting users based on user behaviour	Segment users based on their psyche attributes
	High volume & high dimensional (+1000 entries)	***Wcss + Non-hierarchical Clustering***							
Supervised	**Categorical dependent feature**	***Classification***	Predicting demographic of users	Predicting location of users	Predicting user's time of consumption/Usage of a product or service	Predict user's interest in a brand	Predicting user's probability of purchasing a product/service	Predicting user behaviour profile/a desired behaviour	Predicting users' psyche
	Continuous dependent feature	***Regression***							

Figure 1 – Framework outlining various Strategies to user profiling based on customer attributes and Machine Learning tecniques

Furthermore, solely relying on customer attributes and techniques shown in the figure above, does not yield a profound outcome, but rather often yield simplistic results. For instance, strategies shown in figure 1 for user profiling does not account the nature of user data (Implicit vs Explicit) or segmentation approach (Posteriori vs Apriori). As a result, a model based on the proposed Hybrid user profiling type of Khosrow-Pour (2009) and proposed two-step segmentation approach of Dolnicar (2008) is proposed to address this gap. This model, as

shown in figure 2, outlines various user profiling processes. It illustrates various possibilities of user profiling processes that are complementary to the framework depicted in the figure 1.

		Explicit		Implicit	
		Posteriori	Apriori	Posteriori	Apriori
Explicit	Apriori	E1	E2	Ei1	Ei2
	Posteriori	E3	E4	Ei3	Ei4
Implicit	Apriori	iE1	iE2	i1	i2
	Posteriori	iE3	iE4	i3	i4

Figure 2 Multi-step user profiling process model

In the model shown in figure 2, user data is categorized into two types of Explicit and Implicit as described by Kanoje et al. (2014). To recap, the Explicit data is provided voluntarily by users from website forms or questionnaires and Implicit data are the so-called behavioural data, that is generated when users interact with the website. Moreover, Apriori is also known as the common-sense segmentation, whereas Posteriori is also known as the data-driven segmentation. For instance, in the cell *Ei1*, users are processed first by common-sense groups based on their Explicit data and afterwards, they are segmented on data-driven attributes based on their Implicit data.

2.11. Related research on user profiling

This section of the paper briefly reviews various prior research on user profiling and mentions the techniques and customer attributes used in each of them. Understanding previous work in this area provides an indication as to what has been tried and their results. In general, there are two branches of research of user profiling, one branch is about evaluating and using different techniques and customer attributes and the other branch is about evaluating the performance of various Machine Learning techniques, methods and algorithms. This section, first introduces and critically reviews various researches done regarding the first branch of research, followed by critical review of researchers on performance evaluation of various techniques and their strengths and weaknesses. The end of this section is concluded by a synopsis of previous research on user segmentation.

Many studies evaluated various approaches to behavioural profiling, where some experimented by using one or a mixture of techniques and features. Ahmed, Low, Aly, Josifovski, & Smola (2011) used historical user activity to segment users. They categorised websites into pre-defined categories that lead to the development of dynamic segmentation using the so-called *Topic models* technique that divides webpages into (pre-defined) categories such as *Dating*, *Baseball* (Ahmed et al., 2011). In that study, the researchers made use of Implicit user data and grouped users by applying a probabilistic variation of a clustering technique. In another study, Yao, Eklund, & Back (2010) created customer segments using a two-step technique , namely SOM-Ward clustering technique, based on Demographic customer attributes such as age and shopping behavioural characteristics such as Loyalty points and the spending amount. The aim of their research was to identify high-spending customers and create a prediction model, that can identify such customers, which was successfully done. Furthermore, Zhou & Mobasher (2006) conducted a user profiling research, using a mixture of factor analysers (MFA) technique by employing user's navigational behaviour in browsing sessions. This study aimed to evaluate the performance of using MFA technique based on the shared interest of users and their behavioural observation. The study concluded visitor preference as a latent variable and was able to discover "Heterogenous user segments" from it (Zhou & Mobasher, 2006).

Furthermore, Tsai & Chiu (2004) performed user segmentation based on purchasing behavioural pattern of customers. In this paper, they assumed that customers with similar characteristics (in their case , items purchased) often tend to exhibit similar purchasing behaviour (Tsai & Chiu, 2004). Their approach is quite similar to user segmentation based on Purchased-based customer attributes mentioned earlier in this chapter, thus providing evidence on the success of using such attributes for user profiling. Furthermore, Tsai & Chiu (2004) were able to match user profiles to customers that closely resemble the customers' characteristics.

The aim of this research was to offer customers products or advertisements that are closely associated with their user profile and as subsequent their characteristics.

Another branch of user profiling studies in the scientific literature focuses on methods, which take semantics of various features of behavioural features into consideration (Boratto et al., 2016). Gong, Guo, Zhang, He, & Zhou (2013) and Tu & Lu (2010) have incorporated semantic of user queries in their research, where users are grouped based on the similarity of their semantic queries and click behaviours. These researchers made use of a Probabilistic variation of clustering technique and used two forms of data, namely Implicit such as click behaviour and Explicit such as search queries. By doing so, they demonstrated the possibility of combining two forms of data to form user profiles. Moreover, Wu et al. (2009) also used the semantic technique to segment user behaviour based on search queries. In their researcher, they concluded that using Probabilistic Latent Semantic (PLSUS) for user segmentation can increase the CTR by up to 100% compared to the classical clustering algorithm such as CLUTO and K-means. However, the interpretability of such technique is difficult since users can belong to more than one cluster. As a result, there are more segments compared to the classical clustering techniques. This means that users are exposed to materials which are closely matched to their interest. Subsequently, users probably will end up with a *Filter bubble* as Pariser (2011) phrased it in his book. In addition, the overlapping membership inflates the number of cluster members superfluously, which distorts the user profiling outcome where in reality a user might have either lost interest in a particular interest or have been wrongly led down to a certain path.

In a related study, researchers have found that targeting users with behavioural traits that closely resemble the ads, did not necessarily yield a higher CTR. Rather they observed that higher CTR can be achieved when user's behavioural traits are loosely matched with ads (Lu, Zhao, & Xue, 2016). In other words, the high number of clusters does not necessarily translate to a good solution, specifically when using implicit user data. Therefore, a certain threshold of

between-member homogeneity should be maintained in order to provide diversity in between member. Therefore, it could be hypothesized that creating user profiles that loosely resembles the interest of its members yield a better outcome. In addition, creating user profiles that are loosely coupled to the interest of users could also prevent "Filter bubble". This issue was previously mentioned, where cluster members are only exposed to particular materials, products or services that match to their interest, belief and perspective. Therefore, it could provide an exploring opportunity for profile members to discover new material, without having to feel that they are being closely monitored. As a result, users potentially would feel less self-conscious about their activity and resume their normal behaviour during a website visit.

This chapter provided background for the goal of paper. The definition of *behaviour* in this paper is as Cao (2010) states "activities that present as actions, operations , events or sequences conducted by humans in a specific context and environment in either virtual or physical organization". This is important to define, as the goal in the paper is to develop behavioural profiles of website visitors for higher education. Furthermore, certain processes are required to be applied to the raw data for extracting meaningful insight. Fayyad et al. (1996) called the collection of such processes, *Knowledge Discovery*. They divided *Knowledge Discovery* into two main categories of *Verification* and *Discovery*, where the Discovery category correspond to the two Machine Learning algorithms namely Unsupervised (such as Clustering) and Supervised (such as Classification). Each of the Machine Learning categories has a variety of techniques but this paper focuses on two specific sub-type of each Machine Learning categories, namely *Hierarchical* and *Non-hierarchical* for Unsupervised, and *Classification* and *Regression* for Supervised. Furthermore, Dolnicar (2008) describes three approaches to segmentation, namely Apriori , Posteriori and Hybrid. The first two approach are one-step approach two segmentation, whereas Hybrid is the combination of the two approaches to account for the shortcoming of using each approach on its own. Three types of user profiling

are mentioned in the literature (Kanoje et al., 2014; Khosrow-Pour, 2009). The first one is based on explicit user data such as website forms and the other one is based on implicit user data such as the ways users behave and view websites. Khosrow-pour (2009) proposes the third type, which is based on a combination of the two aforementioned types, and he argues that the accuracy of user profiles is dependent on data quality. Therefore, he states that by combining the two aforementioned types, Hybrid type portrays a more true and realistic image of users in each profile. Moreover, Cufoglu (2014) proposes two methods to user profilin where one, namely *content-based*, focuses on behaviours of a person in the same circumstance and the other one, namely *Collaborative*, focuses on grouping individuals with similar behaviours. Moreover, in order to realize the goal in the paper, the various source of data needs to be realized. A literature review, based on prior research suggests three appropriate sources of data, namely Web data, Business data and Meta data. Using a combination of these sources for creating behavioural profiles could lead to more accurate and richer insights. Lastly, prior research had shown various ways to profiling based on a various combination of methods, approaches, types and customer attributes mentioned in this chapter. However, none provided a framework that describes various ways to user profiling and customer attribute prediction using Machine Learning. Such framework and a complementary model that demonstrates various ways to processes data for user profiling are provided in this chapter. The following chapter describes the methodology used for achieving the goal in the paper.

3. Methodology

This chapter outlines the steps for conducting analysis in order to realize the goal in the paper. It elaborates on methods used, expands on the nature and source of the data, and depicts analysis strategy by means of illustration of necessary steps for the analysis. Following the literature, the main assumption in this paper is that people who manifest similar behavioural pattern, tend to share similar goal or interest. This paper uses static collaborative user profiling method as modifying behavioural profiles are not possible at this stage of research. The goal in the paper is realized by analysing visitors' Behavioural data of the University of Twente to discover behavioural profiles of website visitors and elaborate on their behavioural characteristics as static profiles.

In this research, multiple data sources are used, namely Web data and Business data (CRM data). It is an *exploratory* research that studies the application Machine Learning in behavioural profiling of visitors of a University's website. The goal in the paper is achieved by using the framework and model proposed for User profiling in the previous chapter. The behavioural profiles will be based on data collected from visitors of University of Twente's website. It is a quantitative and yet, exploratory research based on (previously collected thus) existing data to understand typical behaviours UT website visitors and to illustrate the outcome in a simple way sheds a light on behaviours manifested by visitors. The outcome would ideally lead to knowledge discovery of potential prospects and their manifested behavioural patterns.

3.1. Data & Data collection

The data used for this paper is secondary data, meaning data is not collected first hand but rather used an existing database of the University of Twente is used. The data used for this study is the combination of Explicit and Implicit user data. For instance, Web data is an Implicit form of user data (such as behaviour on the website) and interest in a study is an example of Explicit user data. Due sensitivity of such data, the database is handled with care in order to

avoid breaching the privacy of visitors. All data is anonymised and the research is conducted by a sole researcher. In addition, the database (alongside the pre-processed) is processed in an anonymous and careful manner.

Due to fact that data is of secondary nature, the reliability and quality of collected data are unclear although often the collected data via CRM systems and website Cookies are quite reliable. Nonetheless, actions are taken in order to improve the reliability and quality of the database. One way to do so is by cross-checking and validating the data during the pre-processing procedure. Two sources of data are used, namely Web data (Behavioural data) and Business data (CRM database) of the University of Twente. The web data is retrieved from Google analytics' account of the University of Twente.

3.2. Analysis Strategy

This section of the paper outlines how the analysis is conducted in this paper. It provides a clear understanding of how and what steps are necessary for replication with a new dataset. In order to achieve the goal of this paper, analyses are conducted using programs such as R, Python and Microsoft Excel for visualization of outcomes. Using such programs provide freedom for analysis as oppose to SPSS. Additionally, it provides a learning opportunity and possibility to gain experience in R and Python programming languages to the researcher. The chosen technique for conducting analyses of this paper is Unsupervised Machine Learning algorithm called Clustering. This Machine Learning algorithm is used in this paper to unravel the naturally occurring patterns and groups within databased using behavioural features outlined in the following section.

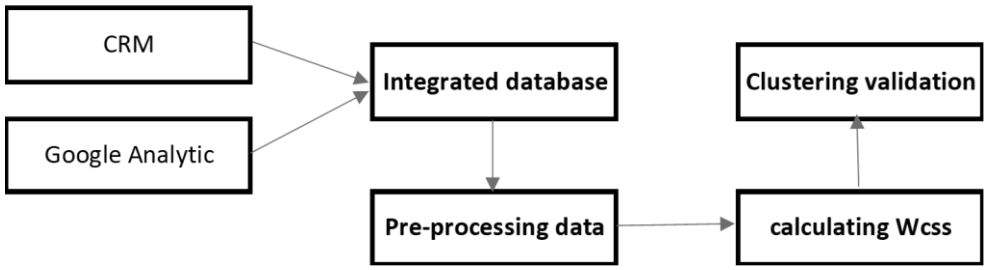

Figure 3 – Illustration of steps for conducting the anlaysis

The figure 3 depicts the overall required steps for conducting each analysis. It starts with retrieving CRM data in form of multiple (excel) documents from Marketing department of the University of Twente and retrieving Web data from Google analytic of the University of Twente. The collected data is for the period between January 1st, 2016 until December 31st, 2017. Once all data is received, they are integrated and transformed into a single cohesive document that has an appropriate format for Machine Learning. The following step is the pre-processing of the integrated file. This step includes processes such as removing invalid data entries, transforming data into a Machine Learning appropriate format and transforming categorical variables into dummy variables.

The pre-processed data of this paper has high volume (about 50,000 data entries), therefore as indicated by the proposed framework in figure 1, a Hierarchical clustering technique is not possible. The reason for it is that such technique requires enormous computational power for high volume data. Therefore, a Non-hierarchical clustering algorithm as suggested by the framework in figure 1, namely K-means is used. The reason why this particular technique (non-probability variation of clustering technique) is chosen, is to avoid superfluously inflate the number of profiles. Furthermore, this technique allows for loose coupling of behaviours and interests with each profile member thus avoiding the *Filter bubble* as well. However, the non-hierarchical (the Non-probabilistic variation of) clustering technique requires Apriori knowledge of the appropriate number of clusters. Since this is not known, the 'within-cluster sum of squares' score (Wcss) is used to determine the appropriate number of profiles.

There are in total four analyses conducted in this paper. Here below, the steps for each analysis is described and the chosen strategies, of the proposed framework and model in this paper, are described. The analysis is done once on all visitors in order to create holistic behavioural profiles of visitors. The aim is to demonstrate the behavioural profiles of all visitors and the typical behavioural pattern of each profile. The steps for conducting this analysis is depicted in figure 4. The first analysis is conducted using only the Implicit data of visitors to

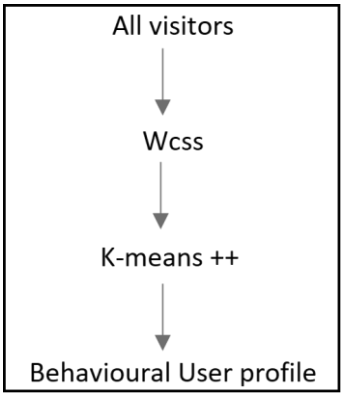

Figure 4 – illustration of steps taken in the 1st analysis

create Posteriori behavioural profiles. The approach in the 1st analysis represents one of the approaches hypothesized by previous researchers in behavioural profiling. It is desirable to see

how behavioural profiles are different when the proposed framework and model introduced in chapter two are used. In addition, it is interesting to see if the discovered behavioural profiles are consistent across various factors such as Study level and Country. Therefore, three more analyses are done using the framework and model proposed in this paper.

The first analysis using the framework and model of in this paper is to control for the country factor. One country is chosen to evaluate the consistency of discovered behavioural profiles compared to behavioural profiles of all visitors. Therefore, the second analysis (analysis controlling for country factor) applies the "Ei1" strategy of *the multi-step processing model* shown in figure 2. Moreover, in this analysis visitors are first grouped by common-sense (Apriori) of Explicit data (country) and then grouped in a Posteriori manner based on their Implicit data, as shown in figure 5. The choice of the country for the second analysis is based on the number of visitors per country, which would preferably be a country with the highest number of visitors since the clustering technique requires large enough sample so that its outcome could be representative of the country.

Furthermore, two more analyses are conducted on the two study levels, namely Bachelor and Master. These two analyses control for study level factors. For conducting analyses on the

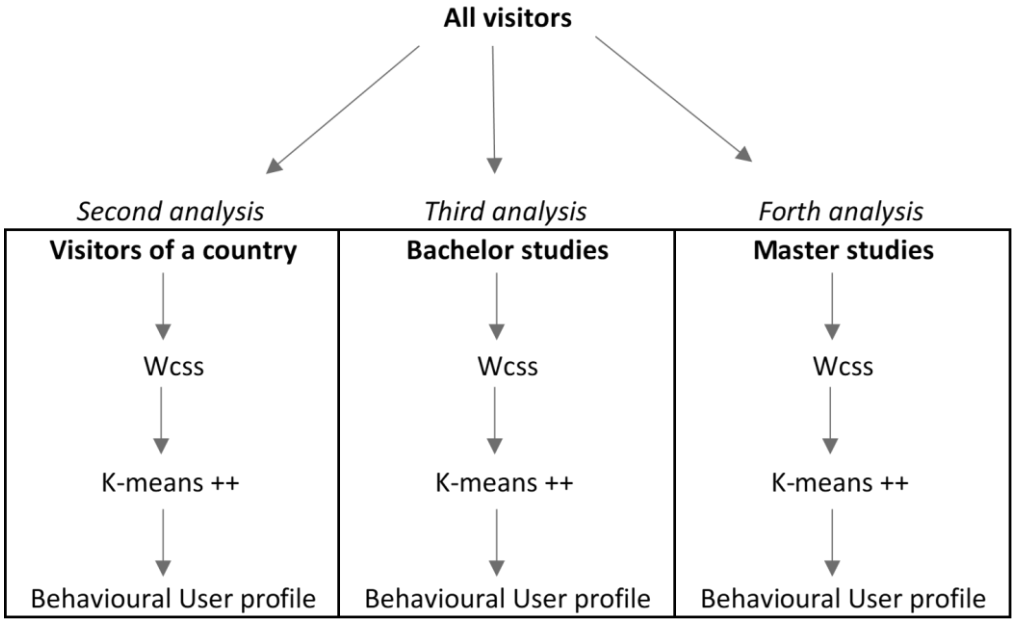

Figure 5 – illustration of steps taken for each analysis

two study levels, the same strategy as in the second analysis, namely "Ei1" of the *multi-step processing model* shown in figure 2 is used. This means that visitors are first grouped by common-sense (Apriori) of Explicit data (study levels) and then grouped based on Implicit data (behavioural data). The outcome of these analyses using the chosen strategy would reveal if the discovered behavioural profiles are consistent compared to the previous analyses. Figure 5 depicts the steps taken for conducting the second, third and the fourth analysis.

The profiling solution of all analyses are described in terms of manifested behaviours of each profile, which is introduced later on in this chapter. In order to evaluate the quality of behavioural profiles, the solutions are validated using cross-validation, silhouette test - homogeneity test and Kruskal-Wallis test to ensure the validity of solutions and by subsequent results concluded from it. the aforementioned tests are introduced and elaborated in the following section.

3.3. Cluster validity

Clustering algorithms always generate clusters but their outcome might not always accurately reflect visitors. As a result, certain tests are conducted to ensure the accuracy, and validity of clustering solutions. Validating the clustering solutions is important as it indicates whether if the visitors and their manifested behaviours are grouped accurately. Review of the scientific literature shows a number of tests and approaches for validation of clustering solutions. However, none proved to perform better than the other. Therefore, the three most frequently used tests in the literature is used for this paper. They are Kruskal-Wallis, Silhouette test (homogeneity test) and cross-validation. Each of these tests and approaches are briefly described in this section.

3.3.1. Silhouette test (Homogeneity test)

Silhouette score (also known as Homogeneity test) is the calculated average distance between each visitor and its cluster centroid (De Amorim & Hennig, 2015; Jain, 2016;

Rousseeuw, 1987). This score indicates how well visitors are distributed within clusters (De Amorim & Hennig, 2015; Rousseeuw, 1987). This score evaluates cohesion of the within-cluster distance of visitors in the same cluster from their cluster centroid (De Amorim & Hennig, 2015).

The outcome of the formula above is a score ranging from -1 to 1, where a score close to -1 means the clustering cohesion is not good and visitors should be readjusted to improve the clustering solution and a score closer to 1 means the clustering solution is quite good and cohesive (De Amorim & Hennig, 2015). The previous researchers concluded that there is no single cluster validity test that has advantage over the other another, but the silhouette score has repeatedly shown to perform well in many research (Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013; De Amorim & Hennig, 2015; Pollard & van der Laan, 2002). Furthermore, Silhouette score can be used for any distance measurement.

3.3.2. Kruskal-Wallis test

Kruskal-Wallis is another validity test that evaluates the validity and accuracy of behavioural profiles discovered. This test provides evidence on the statistical differences between the discovered behavioural profiles. Generally, two set of tests are proposed in the literature depending on the characteristic of data. ANOVA test requires parametric data, and if data is non-parametric then Kruskal-Wallis is suitable (Corder & Foreman, 2009; Solutions, 2017). Either of these tests indicates that at least one cluster statically and significantly is different from the other clusters. However, none indicate on what features such difference exists. Future research could look into between profile difference and conduct post hoc analysis to identify such difference on a granular level. Example of Posthoc analysis is followed by rejection of null-hypothesis for ANOVA test is *Pairwise T-tests* or *Scheffe* or *Bonferroni* whereas for Kruskal-Wallis is *Dunn's test* or *Conover-Iman test* (Conover & Iman, 1979; Dunn, 1964; Williams, 2004).

3.3.3. Cross-validation

Cross-validation is another validation method for Machine Learning. There are 3 variations of Cross-validation, namely *Holdout* , *K-folds* and *Leave-one-out* methods (Schneider, 1997). This validation test focuses on evaluating if the correct number of profiles are used for clustering visitors. The researcher of this paper is interested to use the *Holdout* method. In this version of Cross-validation method, the original dataset is randomly split into two sub-samples with two sub-samples having a ratio of 75% (Test dataset) to 25% (Train dataset) of the original dataset. The validation of clustering solution using Holdout is done by comparing the recalculated Wcss score and the number of clusters for each sub-sample (Test dataset and Train dataset). The judgement on performance of the two randomly selected sub-sample is based on the indicated number of clusters by looking at the Wcss scores.

3.4. Behavioural Features

Pre-processing the raw data and transforming it a useable Machine Learning format,21 behavioural features and one explanatory feature was extracted from the pre-processed data. In order to improve the interpretability of the chosen technique (cluster analysis), an explanatory feature is included. This explanatory feature provides insight about the discovered behavioural profiles of website visitors and provides an indication which of the behavioural profiles have the highest conversion rate. As previously mentioned, in this research the conversion point is the “Eligibility test”, which will be explained more in the following section.

3.4.1. Explanatory variable

The aim of this study is to find out behavioural profiles of website visitors of UT. It is known to the Marketing and Communication department of the University of Twente that *Eligibility Check* is one the last behaviours that is manifested in the registration process. Therefore, it is desirable to discover behavioural profiles that generate insight about visitors who take taking such test. This is important as future research could create prediction model

based on manifested behaviours of desired behavioural profiles in order to focus marketing activities to such visitors with appropriate advertisement message through appropriate channels. In this manner, the marketing activities can be done more effectively and efficiently.

As a consequence of effective and efficient marketing activities, University of Twente is able to attract more visitors to take the Eligibility Check, thus collects detailed personal data from its visitors. Such data is can be used to identify the common weaknesses of potential applicants. Such insight could be used to offer courses to applicants in order to compensate for their skill/knowledge deficiency.

3.4.2. Behavioural and Behavioural Source Features

The pre-processing of the raw data generated many variables, which could potentially explain why visitors do or don't do the Eligibility check. The extracted behavioural features from the pre-processed data are divided into two main categories. The first category is the type of online behaviours manifested by visitors and the second category indicate how visitors found their way into the University's website to manifest behaviours of the former category. The first category of features (from now on will be referred to as 'Behavioural features') are as follow:

(Educational) Brochure request	PDF download
Request student for a day	Managed CTA Click
Question via web form	Managed CTA Display
Fair attendance	Scholarship Finder
Open day registration	Frequently asked questions

The first feature, *Brochure request* is when visitors download at least one educational brochure. Next feature, *Request student for a day*, is when a visitor registers to try out a day with a student at the university. *Question via web form* is when a visitor asks a question via the

web form available in the web pages. Following feature, *Fair attendance*, is when a visitor attended at least one of the fairs of University of Twente. The next feature, *Open day registration*, is when a visitor registers for open days at the university to get information about the study of interest. The next feature, *PDF download*, is when a visitor downloads a brochure that is non-study related. For instance, additional information such as finance and catalogue of the University. The *Managed CTA* (Click or Display) is when a visitor returns to the website and therefore, the call to action message and content changes to encourage the visitor to take an action. Often times such actions include registering for an event or even sending the application. *Scholarship finder* is the behavioural that a visitor manifest when he/she goes to the scholarship webpage and look for scholarship availability. The last feature is behaviour of visitors when they seek information from the *Frequently asked question*.

The second category of features (from now on will be referred to as 'Behavioural source features') are as follow:

Direct	Outlook
Google	Quick link
Facebook	Master portal
Gmail	Mail invitation
Bing	Program route mail

Each one of the above-mentioned sources of behavioural features indicates where visitors entered the university's webpage to manifest the behaviours described earlier. Creating such distinction among various extracted features from the original data is an important part of discovering similarity among visitors otherwise stated, their behavioural patterns. The first category of features reveals the important behavioural attributes and patterns of groups and the

second category of features reveal the important entry points for visitors who exhibit certain behaviours.

4. Results

In this chapter, the proposed framework and model are applied to website visitors of the University of Twente. By doing so, the goal of this paper is realized and evidence of added value by using the framework and data processing model proposed in this paper is provided. The analysis outcomes on behavioural visitor's data in this chapter are demonstrated in forms of tables and figures. In addition, the discovered behavioural profiles among visitors of UT website are described.

4.1. Data Description

After the pre-processing data, data contained about 49,110 visitors. As shown in table 1, there is no missing value for any chosen feature. The range in the pre-processed database equals to one. This is simply due to the fact that all categorical features used in this paper are transformed to dummy variables. Each dummy feature in the table below indicates the presence

Table 1

Descriptive Statistic of pre-processed database

Features	N	Range	Minimum	Maximum	Mean
Brochurerequest	49110	1	0	1	0.28735
Registerd student for a day	49110	1	0	1	0.01354
Question via webform	49110	1	0	1	0.09324
Fair attended	49110	1	0	1	0.01128
Oped day registration	49110	1	0	1	0.00098
PDF download	49110	1	0	1	0.00183
Managed CTA Click	49110	1	0	1	0.00301
Managed CTA Display	49110	1	0	1	0.01004
Scholarship finder	49110	1	0	1	0.00108
Frequently asked Question	49110	1	0	1	0.00055
Source Direct	49110	1	0	1	0.00161
Source Google	49110	1	0	1	0.00935
source Facebook	49110	1	0	1	0.00049
Source Gmail	49110	1	0	1	0.00039
Source Bing	49110	1	0	1	0.00043
Source Outlook	49110	1	0	1	0.00041
Source Quicklink	49110	1	0	1	0.00092
Source Master portal	49110	1	0	1	0.00041
Source Uitnodigingmai	49110	1	0	1	0.00016
Source Programma route mail	49110	1	0	1	0.00012
Eligibility check	49110	1	0	1	0.73335
Valid(N)	49110				

or absence of a behaviour (1 for presence and 0 for the absence of it). Moreover, the minimum and maximum might not be necessary a meaningful descriptive statistic for such features, but it provides a good indication if there is an error in the data. Table 1 shows no irregularity or error within the database when evaluating range, minimum and maximum, considering the nature of features (categorical transformed to dummy variables).

The mean of each feature shown in table 1, represents the frequency proportion of each feature. For instance, approximately 28.7% of visitors requested some form of Educational brochure. In addition, nearly 2% of visitors registered to be a student for a day, 9% asked a Question via web form. In total, 73% of the visitors took the Eligibility check. About 1% of visitors attended a Fair and about 1% of visitors entered via Google. Other features seemingly have frequency proportion of less than 1%, which might seem too few but even such small proportion translates to roughly 490 visitors.

The data size is large enough to fulfil the pre-requisition for cluster analysis (having 4 to 5 times observation as features). In addition, there are not any outliers or irregularities in the data. The data does not require any standardization, as categorical features are already transformed into dummy variables where each feature indicates the presence or absence of a behaviour. Therefore, range and distance of all features are identical. As result, the data is ready for analysis and the first step in all analyses as described in chapter 3 is calculating Wcss score to determine the number of profiles for each analysis. The following section describes the outcome of Wcss score and number of profiles for each analysis.

4.2. Calculating Wcss

Often in the literature, the combination of cluster analysis methods (two-step clustering using hierarchical and non-hierarchical) is used to reach a valid result by using hierarchical clustering to determine the number of clusters and non-hierarchical to assign cases to cluster(s). However, as mentioned in model introduced in chapter 3, the large volume of data prepared for this paper does not allow for hierarchical clustering. As a result, a Non-hierarchical clustering technique is an appropriate choice. However, the number of profiles needs to be determined, which is done by calculating 'within-cluster sum of squares' (Wcss) score. Figure 7 demonstrates the Wcss score against number of profiles for each analysis.

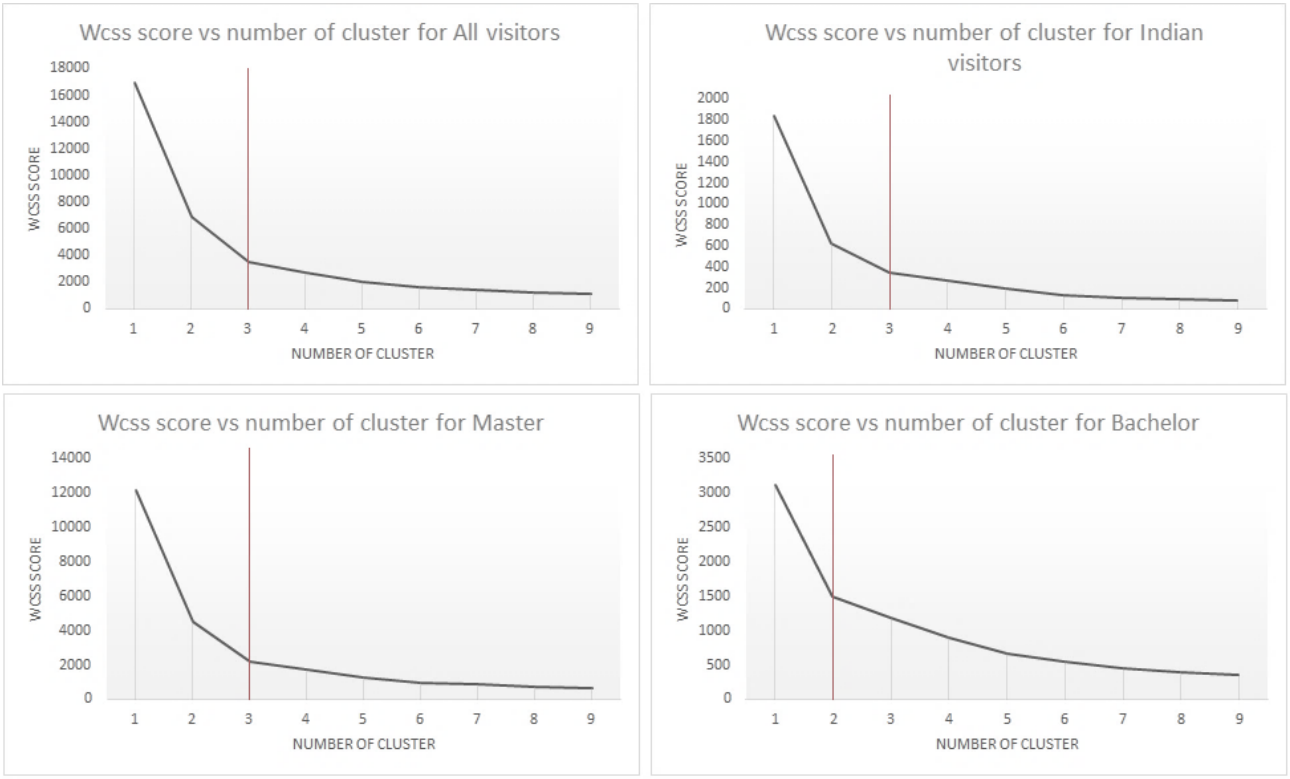

Figure 6 – Wcss score scree plot of all analysis

The appropriate number of profiles in graphs above are the points where Wcss score doesn't change dramatically or in another word the latest angled dip in each graph. These points are the vertical lines shown in each graph. This method of determining the appropriate number of profiles based on the graphical visualization of Wcss is referred to as Elbow method.

The appropriate number of profiles using this method for All visitors, Indian visitors and visitors interested in Master are three. However, the appropriate number of profiles for visitors interested in Bachelor seems to be two. Here the controlling country is determined India. The reasoning behind this choice is elaborated in the section 4.3.2. Knowing the appropriate number of profiles for each analysis, the Non-hierarchical variation of clustering analysis can be conducted and the outcome of such analyses are described in the following section.

4.3. Cluster analysis

Following the methodology chapter, the Behavioural profiling (clustering analysis) is performed four times. The First Behavioural profiling is on all visitors regardless of the country of origin or study levels. The second Behavioural profiling is performed only on Indian visitors to see whether the discovered pattern is independent of country. Lastly, the Behavioural profiling is performed on each study levels to see whether if the discovered behavioural profiles are consistent across study levels, in order word, if they are independent of study level as well.

4.3.1. Behaviour profiling of all visitors

The analysis using K-means (non-hierarchical variation used here is K-means ++) clustering method using three 3 seeds (determined in the 4.2 section) and maximum iteration of 300, yields the result as shown in table 2. There are 9.3% of visitors in the first cluster, 64 % in the second and 26.7 % in the third cluster. There is not any unassigned visitor to clusters. From here on in this paper, the terms cluster and profile are used interchangeably. Table 3 demonstrates the clustering outcome where all clusters are represented in terms of frequency of previously-defined behavioural features manifested by each cluster members.

Table 2

Visitor distribution in each cluster

	N	%	Valid %
Cluster 1	4579	9.32	9.32
Cluster 2	31422	63.98	63.98
Cluster 3	13109	26.69	26.69
Total	49110	100.00	100.00

In the 1st cluster, about 41.3% visitors converted (the measure for conversion is people who took the Eligibility check). All cluster members in the 1st cluster (100%) asked at least once a Question via web forms, 21.9% downloaded some form of "Educational brochure" (educational brochures provide study specific information) and 4.2% requested to be "a student for a day". It can be hypothesized that visitors in this cluster are either curious about programs offered by the University of Twente or they require additional information that was not available on University's website. This group of visitors seem to be interested to study at UT

The members of the 1st cluster manifest behaviours that indicate their interest in studying at UT. They look through available information on the website or other sources but did not find what they seek. As result, they manifest behaviour such as Question via website to obtain more information. Such characteristics resemble the interest phase of AIDA model, where audience's interest is piqued. This phase is associated with customers who would like to acquire enough knowledge and at the same time, they have developed affiliation for the institute to some degrees (Wijaya, 2012). Therefore, this cluster will be called the *Interested* group.

The majority of visitors in the 1st cluster come from Netherlands, India and Germany and the most popular studies are Psychology (accounts for 8.28% of the 1st cluster), Business Administration (accounts for 6.94% of the 1st cluster) and Mechanical engineering (accounts for 6.4% of the 1st cluster).The comprehensive statistic about the country of origin and popular studies of each cluster are depicted in table 4 and 5. It is worth noting that in this analysis no distinction is made between the two study types (master or bachelor) and only focused on the overall behaviour of visitors.

The 2nd cluster as shown in table 3, comprise of visitors of who nearly 99.7% took Eligibility check and about 1% attended Fair. This cluster represents the majority of UT's website visitor, which translate to nearly 64% of total visitors. The most frequently manifested behavioural features in this cluster are Brochure Request or Question via Web form. This

indicates that members of this cluster had acquired the information they wanted either by looking at the information available on the University's website or other external sources. Visitors in this group manifest behaviours which could be hypothesized as the same as people who are in the Desire phase of the AIDA model. In this phase of the AIDA model, the audience has developed a favourable attitude towards a brand, here in this paper the Institute, thus would like to know if it is possible for them to study in UT (Lewis, 1908; Rawal, 2013) . As a result, visitors go to the University's website to take the Eligibility check. Thus, this group is called *Desirability* profile.

Table 5 demonstrates that the top three popular studies in the Desirability profile are Business Administration, Computer science and Mechanical engineering. The top three visitor countries in the 2nd cluster are Non-EU countries such as India, Nigeria and Pakistan. This is an interesting finding as it indicates visitors outside Europe, speciality in Africa and Asia, visitors would like to know if their educational background is sufficient for the University of Twente. Furthermore, it is interesting to evaluate behaviours of Indian visitors and compare it to the behavioural profile of all visitors. Indian visitor represents the largest group from all countries as well as in the second cluster.

The last cluster, namely the third cluster, comprised of visitors whom all downloaded Educational brochure as shown in table 3. In terms of size, this cluster represents the second biggest group of all, containing nearly 13,000 visitors. This group much like the Desirability group, its members never asked any Question via web form. However, this group seem to be different than the other two clusters in the way that they are more influenced by Managed CTA (in both Display and Click). In addition, they sought information more often via FAQ (frequently asked questions).

It seems as if visitors in this cluster seek to know more and acquire additional information about their desired studies by downloading Educational brochures. This is a

distinctive characteristic of this group as all of the visitors in this group download Educational brochures. It could be hypothesized that visitors in this cluster came across UT advertisements and would like to know more about programs details. Therefore, they have downloaded various form of Educational brochures. The manifested visitor behaviour of this cluster is indicative of their stage within the AIDA model. Potentially it could be hypothesized that such visitors are in the Attention stage of AIDA model. In this stage, customers become aware of the service or product and seek to inform themselves. Especially considering that this stage of AIDA model is associated with cognition and rational knowledge seeking (Lewis, 1908; Rawal, 2013) .As result, this group is called the *Attention* group in this paper.

The table 4 shows that the top visitor country in Attention group is an EU country, namely the Netherlands followed by two Non-EU countries namely, India and Indonesia. Moreover, the top two popular studies in this cluster are Mechanical engineering and Electrical Engineering, as shown in table 5. This group is dissimilar from the other two clusters in terms of visitor's interest in studies.

Table 3

Distribution of behavioural features of all visitors in each cluster

	Cluster 1- Interest		Cluster 2 - Desireability		Cluster 3 - Attention	
Behavioural categories	N	%	N	%	N	%
Behavioural features						
Brochurerequest	1003	21.904	0	0.000	13109	100.000
Request student for a day	191	4.171	174	0.554	300	2.289
Question via webfrom	4579	100.000	0	0.000	0	0.000
Fair	70	1.529	313	0.996	171	1.304
Open day registration	6	0.131	14	0.045	28	0.214
Pdf download	24	0.524	12	0.038	54	0.412
Managed CTA Click	30	0.655	12	0.038	106	0.809
Managed CTA Display	80	1.747	46	0.146	367	2.800
Scholarship finder	13	0.284	1	0.003	39	0.298
Frequently asked question	11	0.240	5	0.016	11	0.084
Eligibility check	1892	41.319	31339	99.736	2784	21.237
Behavioural source features						
Source Direct	10	0.218	6	0.019	63	0.481
Source Google	87	1.900	51	0.162	321	2.449
Source Facebook	4	0.087	2	0.006	18	0.137
Source Gmail	1	0.022	2	0.006	16	0.122
Source Bing	0	0.000	4	0.013	17	0.130
Source Outlook	8	0.175	5	0.016	7	0.053
Source Quicklink	7	0.153	2	0.006	36	0.275
Source Master portal	3	0.066	1	0.003	16	0.122
Source Uitnodigingmai	1	0.022	2	0.006	5	0.038
Source Programma route mail	2	0.044	1	0.003	3	0.023

Table 4

Distribution of visitor countries in each cluster

Cluster 1 - Interest			Cluster 2 - Desireability			Cluster 3 - Attention		
Country	N	% Country		N	% Country		N	%
			Asia					
India	378	8.255	India	3981	12.669	India	1588	12.114
Indonesia	156	3.407	Pakistan	1763	5.611	Indonesia	688	5.248
Pakistan	145	3.167	Indonesia	1507	4.796	Pakistan	364	2.777
Turkey	106	2.315	Iran	898	2.858	Turkey	200	1.526
Iran	102	2.228						
China	101	2.206						
			Africa					
			Nigeria	1815	5.776	Nigeria	424	3.234
			Ghana	1402	4.462	Ghana	309	2.357
			Europe					
Netherlands	1008	22.014	Netherlands	1539	4.898	Netherlands	3227	24.617
Germany	343	7.491	Germany	1306	4.156	Germany	631	4.813
United Kingdom	100	2.184				United Kingdom	203	1.549
			America					
United States	115	2.511	United States	1011	3.217	United States	295	2.250
			Brazil	929	2.957			

Table 5

Distribution of popular studies in each cluster

Cluster 1 - Interest			Cluster 2 - Desireability			Cluster 3 - Attention		
Studies	N	%	Studies	N	%	Studies	N	%
Psychology	379	8.277	Business Administration	3373	10.735	Mechanical Engineering	862	6.576
Business Administration	318	6.945	Computer Science	2414	7.683	Electrical Engineering	704	5.370
Mechanical Engineering	293	6.399	Mechanical Engineering	2219	7.062	Civil Engineering and Management	624	4.760
Electrical Engineering	288	6.290	Civil Engineering and Management	1778	5.658	Psychology	548	4.180
Sustainable Energy Technology	202	4.411	Sustainable Energy Technology	1744	5.550	Sustainable Energy Technology	546	4.165
Industrial Engineering and Management	167	3.647	Electrical Engineering	1735	5.522	Industrial Engineering and Management	534	4.074
Computer Science	162	3.538	Communication Studies	1702	5.417	Environmental and Energy Management	497	3.791
Biomedical Engineering	161	3.516	Industrial Engineering and Management	1631	5.191	Geo-information Science and Earth Observation	458	3.494
Communication Studies	161	3.516	Business Information Technology	1424	4.532	Health Sciences	434	3.311
International Business Administration	147	3.210	Biomedical Engineering	1404	4.468	Business Administration	424	3.234

4.3.2. Behaviour profiling of Indian visitors

In order to see if the discovered behavioural profiles are indeed country independent, the cluster analysis is conducted on a country. As indicated in the previous section, India is chosen as it is an interesting country since there are many visitors from this country. Additionally, the analysis in the previous section indicated that Indian visitors are the largest group in the Desirability group. There are in total 5947 visitors from India, as shown in table 6. The proportion of visitor distribution among clusters remains approximately the same in this analysis as in the previous analysis. The 1st cluster in this analysis compromise of nearly 70% of visitors, the 2nd cluster about 28 % and the remaining 5% visitors belong to the 3rd cluster.

Table 6

Visitor distribution in each cluster

	N	%	Valid %
Cluster 1	3981	66.941	66.941
Cluster 2	1674	28.149	28.149
Cluster 3	292	4.910	4.910
Total	5947	100.000	100.000

The table 7, illustrate the clustering solution of Indian visitors based on the two behavioural categories defined in chapter 3. Comparing Indian visitors with all visitors in terms of behavioural features, Indian visitors have approximately the same three behavioural profiles. Much like the three-unravelled behavioural profiles in the previous analysis on all visitors, Indian visitors have the same three behavioural profiles. The members of the 3rd cluster as shown in table 7, are visitors who manifest Question via web form behaviour and were called Interested group in the previous analysis. Furthermore, the members in the 1st cluster as shown

in table 7, are visitors who did not manifest Request Brochure nor Question via web form behaviour but they all converted (the conversion point is the same as before, taking the eligibility test) and were called the Desirability group in the previous analysis. Finally, the members of the 2nd cluster in table 7 are visitors whom all manifested Educational Brochure behaviour, but unlike the members of the Attention group, only some visitors manifested Question via web form behaviour. Nonetheless, due to high resemblance, this group is also named the Attention group.

Table 7

Distribution of behavioural features of Indian visitors in each cluster

	Cluster 1 - Desireability		Cluster 2 - Attention		Cluster 3 - Interest	
Behavioural categories	N	%	N	%	N	%
Behavioural features						
Brochurerequest	0	0.000	1674	100.000	0	0.000
Request student for a day	4	0.100	6	0.358	0	0.000
Qquestion via webfrom	0	0.000	86	5.137	292	100.000
Fair	74	1.859	26	1.553	6	2.055
Open day registration	0	0.000	0	0.000	0	0.000
Pdf download	1	0.025	12	0.717	0	0.000
Managed CTA Click	0	0.000	17	1.016	1	0.342
Managed CTA Display	2	0.050	47	2.808	2	0.685
Scholarship finder	0	0.000	11	0.657	0	0.000
Frequently asked question	0	0.000	3	0.179	0	0.000
Eligibility check	3981	100.000	546	32.616	179	61.301
Behavioural source feature						
Source Direct	0	0.000	0	0.000	1	0.342
Source google	3	0.075	39	2.330	2	0.685
Source facebook	0	0.000	2	0.119	0	0.000
Source gmail	0	0.000	4	0.239	0	0.000
Source Bing	0	0.000	3	0.179	0	0.000
Source outlook	1	0.025	0	0.000	0	0.000
Source quicklink	0	0.000	10	0.597	0	0.000
Source Masterportal	0	0.000	3	0.179	0	0.000
Source Uitnodigingmai	0	0.000	0	0.000	0	0.000
Source Programma route mail	0	0.000	0	0.000	0	0.000

The clustering analysis on Indian visitors indicates that the behavioural profiles discovered from all visitors are country independent. This is clearly visible when looking at the

prominent behaviours manifested by Indian visitors in each cluster. There are some differences in certain behaviours manifested by each group, especially in number of behaviours that are infrequently manifested. Such differences are discussed in depth in the next chapter.

4.3.3. Behaviour profiling on study levels

In this section of the paper, cluster analysis is performed on each of the two study levels. The two groups are divided (Apriori) by visitors' interest in study levels, namely master and bachelor. As described in chapter 3, the purpose of conducting such analyses on different study levels is to find out whether if the discovered behavioural profiles are consistent across study levels as well. The first sub-section is the outcome of clustering analysis on visitors who are interested in Bachelor studies and the following sub-section focuses on analysing visitors who are interested in Master studies.

Bachelor

The clustering solution based on the appropriate number of profiles indicated in section 4.2, generates the results shown in tables below. The distribution of visitors interested in Bachelor studies among 2 profiles are illustrated in the table 8. The 1st cluster comprise of nearly 31% and the 2nd cluster comprise of 69% of visitors interested in Bachelor studies.

Table 8

Visitor Distribution in each cluster

	N	%	Valid %
Cluster 1	1739	30.993	30.993
Cluster 2	3872	69.007	69.007
Total	5611	100.000	100.000

The detailed outcome of clustering solution for visitors interested in bachelor studies is demonstrated in table 9. The outcome of clustering analysis on visitors interested in Bachelor studies indicates that the discovered behavioural profiles resemble the discovered behavioural profiles of the previous sections with one exception. The two behavioural profiles discovered among visitors interested in bachelor studies resembles the behavioural patterns discovered

previously, namely the Attention group, where majority (almost all) visitors downloaded some form of Educational brochure, and the Interested Group, where the majority of visitors asked Question via web form. However, as shown in table 9, it seems that the Desirability group does not exist among visitors who are interested in Bachelor studies.

Table 9

Distribution of behavioural features of bachelor studies in each cluster

	Cluster 1 - Interest		Cluster 2 - Attention	
Behavioural categories	N	%	N	%
Behavioural features				
Brochurerequest	393	22.599	3872	100.000
Request student fo	237	13.629	209	5.398
Qquestion via webf	1511	86.889	0	0.000
Fair	32	1.840	36	0.930
Open day registrati	4	0.230	20	0.517
Pdf download	9	0.518	18	0.465
Managed CTA Click	10	0.575	42	1.085
Managed CTA Displ	39	2.243	136	3.512
Scholarship finder	0	0.000	9	0.232
Frequently asked q	4	0.230	5	0.129
Eligibility check	206	11.846	63	1.627
Behavioural source feature				
Source Direct	4	0.230	21	0.542
Source google	47	2.703	124	3.202
Source facebook	1	0.058	8	0.207
Source gmail	1	0.058	7	0.181
Source Bing	3	0.173	5	0.129
Source outlook	2	0.115	1	0.026
Source quicklink	1	0.058	6	0.155
Source Masterporta	1	0.058	1	0.026
Source Uitnodiging	0	0.000	3	0.077
Source Programma	2	0.115	3	0.077

Further investigation on the University's website showed that the absence of Desirability group within visitors interested in bachelor studies is due to lack of (at least not an explicit one) Eligibility check for visitors interested in bachelor studies. This explains as to why the third group (the Desirability group) is missing in clustering solution of visitors interested in Bachelor studies. Yet, the table 9 indicates that small proportion of visitors in each cluster have done the Eligibility check. Logically (since there is no explicit Eligibility check available for

bachelor studies), it can be assumed that those visitors went to pages of Master studies, where Eligibility check is available and have done Eligibility check presumably there.

The descriptive statistic of visitors' countries and popular studies among visitors interested in Bachelor studies is shown in tables table 10 and 11. In the 1st cluster, as shown in table 10, the top visitor's countries are from Netherlands, Germany, United States and the United Kingdom. Furthermore, the top three popular studies in this cluster are Psychology, International Business Administrations and Technical Computer science. In the 2nd cluster, visitor's countries are slightly different compared to the 1st cluster. In this cluster, India a non-EU country, took the third place wherein the 1st cluster the United Kingdom is on the third place. Nevertheless, Germany and The Netherlands are the top two countries in both clusters. Moreover, the two identified clusters differ in terms of popular studies as well as the proportion distribution. In both clusters, Psychology is the most popular studies. However, comparing the second most popular studies, it seems Electrical engineering is more popular in the 2nd cluster, wherein the 1st cluster the second popular study is International business administration. One possible explanation for such difference might be that Indian visitors are more interested in Electrical engineering where UK visitors are more interested in International Business Administration.

Table 10

Distribution of visitors interested in bachelor studies by cluster

	Cluster 1 - Interest			Cluster 2 - Attention	
Country	N	%	Country	N	%
		Asia			
Turkey	35	2.013	India	224	5.785
India	33	1.898	Pakistan	89	2.299
Pakistan	28	1.610	Indonesia	88	2.273
			Turkey	71	1.834
		Africa			
Egypt	23	1.323			
		Europe			
Netherlands	670	38.528	Netherlands	1136	29.339
Germany	234	13.456	Germany	409	10.563
United Kingdom	44	2.530	Italy	63	1.627
Spain	24	1.380			
Romania	22	1.265			
		America			
United States	36	2.070	United States	117	3.022

Based on the finding of this section, it can be hypothesised that the discovered behavioural profiles remain consistent in Bachelor studies as well, with an exception that it lacks Desirability group. However, in order to make sure the discovered behavioural profiles are consistent, one last analysis is performed on visitors interested in Master studies in the following section.

Table 11

Distribution of visitors interested in bachelor studies by cluster

Cluster 1 - Interest			Cluster 2 - Attention		
Studies	N	%	Studies	N	%
Psychology	290	16.676	Psychology	421	10.873
International Business Administration	154	8.856	Electrical Engineering	407	10.511
Technical Computer Science	124	7.131	Technical Computer Science	326	8.419
Mechanical Engineering	121	6.958	Mechanical Engineering	296	7.645
Electrical Engineering	121	6.958	International Business Administration	292	7.541
Civil Engineering	88	5.060	University College Twente (ATLAS)	288	7.438
Business & IT	87	5.003	Creative Technology	272	7.025
Creative Technology	78	4.485	Industrial Design	232	5.992
Industrial Design	73	4.198	Civil Engineering	206	5.320
Biomedical Technology	70	4.025	Business & IT	173	4.468

Master

In this section, clustering solution on visitors interested in Master studies is described. The aim of it is to see whether if the discovered Behavioural profiles in previous sections are consistent when controlling for Master studies. The number of profiles was determined in section 4.2. The outcome of the analysis is shown table 12 demonstrate that the second cluster comprises of 70%, representing the biggest group in this analysis, the first cluster comprises of nearly 22% of visitors and Lastly, the third cluster comprises of about 6 % visitors. In terms of visitor distribution among the three pre-defined clusters, the clustering outcome is approximately the same as all visitors and Indian visitors.

Table 12

Distribution of master study level visitors in each cluster

	N	%	Valid %
Cluster 1	9847	22.637	22.637
Cluster 2	31194	71.712	71.712
Cluster 3	2458	5.651	5.651
Total	43499	100.000	100.000

Furthermore, as shown in table 13, the distribution of behavioural features of each cluster indicates that the behavioural profiles unravelled in this analysis resemble the behavioural profiles discovered in previous sections of this report. The three behavioural profiles discovered in previous sections, namely *Attention* group and *Desirability* group and *Interest* group, correspond to cluster 1, 2 and 3 of master studies' clustering solution in that order. As result, one can presume that the behavioural profiles are consistent when controlling for visitors interested in Master studies.

Table 13

Distribution of behavioural features of Master studies in each cluster

	Cluster 1 - Attention		Cluster 2 - Desireability		Cluster 3 - Interest	
Behavioural categories	N	%	N	%	N	%
Behavioural features						
Brochurerequest	9847	100.000	0	0.000	0	0.000
Request student for a day	111	1.127	78	0.250	30	1.221
Qquestion via webfrom	610	6.195	0	0.000	2458	100.000
Fair	152	1.544	299	0.959	35	1.424
Open day registration	9	0.091	12	0.038	3	0.122
Pdf download	51	0.518	11	0.035	1	0.041
Managed CTA Click	83	0.843	8	0.026	5	0.203
Managed CTA Display	280	2.844	28	0.090	10	0.407
Scholarship finder	42	0.427	1	0.003	1	0.041
Frequently asked question	14	0.142	4	0.013	0	0.000
Eligibility check	3135	31.837	31191	99.990	1420	57.771
Behavioural source feature						
Source Direct	46	0.467	6	0.019	2	0.081
Source google	244	2.478	32	0.103	12	0.488
Source facebook	13	0.132	1	0.003	1	0.041
Source gmail	10	0.102	1	0.003	0	0.000
Source Bing	12	0.122	1	0.003	0	0.000
Source outlook	13	0.132	3	0.010	1	0.041
Source quicklink	36	0.366	2	0.006	0	0.000
Source Masterportal	18	0.183	0	0.000	0	0.000
Source Uitnodigingmai	3	0.030	2	0.006	0	0.000
Source Programma route mail	1	0.010	0	0.000	0	0.000

Among all behavioural profiles, Mechanical engineering remains the most popular study. Business Administrations is the most popular study among visitors in Desirability (2nd cluster) and Interest group (3rd Cluster). In contrast, the most popular studies in the Attention group (1st cluster) is Civil engineering and Management. Moreover, the visitor country of the Desirability group (2nd cluster) are mostly Non-European countries such as India, Nigeria, Pakistan and Indonesia. As result, it could be hypothesized that Asian and African countries are more likely to belong to Desirability group and thus be more interested in studies such as

Business Administration, Civil Engineering and Mechanical Engineering. Moreover, the outcomes suggest that Dutch visitors are more likely to be interested in Civil Engineering studies where German visitors are more likely to be interested in Business Administration. The tables 14 and 15 illustrate the popular studies and country of visitors in each profile.

Table 14

Distribution of visitors interested in Master studies by cluster

Cluster 1 - Attention			Cluster 2 - Desireability			Cluster 3 - Interest		
Country	N	%	Country	N	%	Country	N	%
			Asia					
India	1440	14.624	India	3978	12.752	India	272	11.066
Indonesia	635	6.449	Pakistan	1763	5.652	Indonesia	106	4.312
Pakistan	294	2.986	Indonesia	1506	4.828	Pakistan	98	3.987
			Iran	898	2.879	Iran	80	3.255
						Turkey	67	2.726
			Africa					
Nigeria	396	4.022	Nigeria	1815				
Ghana	276	2.803	Ghana	1401				
Ethiopia	162	1.645						
			Europe					
Netherlands	2197	22.311	Netherlands	1398	4.482	Netherlands	373	15.175
Germany	254	2.579	Germany	1260	4.039	Germany	123	5.004
			Italy			Greece	67	2.726
			America					
United States	190	1.930	United States	1010	3.238	United States	68	2.766
Brazil	163	1.655	Brazil	929	2.978			

Table 15

Distribution of visitors interested in master studies by cluster

Cluster 1 - *Attention*			Cluster 2 - *Desireability*			Cluster 3 - *Interest*		
Studies	N	%	Studies	N	%	Studies	N	%
Civil Engineering and Management	653	6.631	Business Administration	3345	10.723	Business Administration	276	11.229
Sustainable Energy Technology	613	6.225	Computer Science	2409	7.723	Mechanical Engineering	154	6.265
Mechanical Engineering	599	6.083	Mechanical Engineering	2204	7.065	Electrical Engineering	150	6.103
Environmental and Energy Management	502	5.098	Civil Engineering and Management	1771	5.677	Biomedical Engineering	141	5.736
Geo-information Science and Earth Observation	459	4.661	Sustainable Energy Technology	1741	5.581	Computer Science	134	5.452
Industrial Engineering and Management	455	4.621	Electrical Engineering	1725	5.530	Sustainable Energy Technology	130	5.289
Business Administration	446	4.529	Communication Studies	1686	5.405	Communication Studies	121	4.923
Construction Management and Engineering	416	4.225	Industrial Engineering and Management	1620	5.193	Psychology	112	4.557
Master Risk management	412	4.184	Business Information Technology	1422	4.559	Civil Engineering and Management	105	4.272
Spatial Engineering	385	3.910	Biomedical Engineering	1395	4.472	Industrial Engineering and Management	99	4.028

4.3.4. Interpretation of analyses

Previous sections of this chapter indicated that there three Behavioural profiles among website visitors of the University of Twente. They are *Attention*, *Interest* and *Desirability* profiles. Each profile is comprised of a distinctive behavioural pattern where each profile is described in terms of its prominent features. However, focusing on prominent behavioural features might not be indicative of a true and realistic picture of profiles. Therefore, this section aims to evaluate and assess findings of previous sections in depth to uncover details that might be only visible when findings are compared in a side-by-side manner. This section begins with comparing visitor's distribution to each profile for each of analyses followed by a comparison of profiles of all analyses in terms of behavioural features and their patterns. Next, a brief description of the Behavioural source, which is basically the entry point of visitors to UT's website, is given. Finally, this section ends with a brief summary of this section.

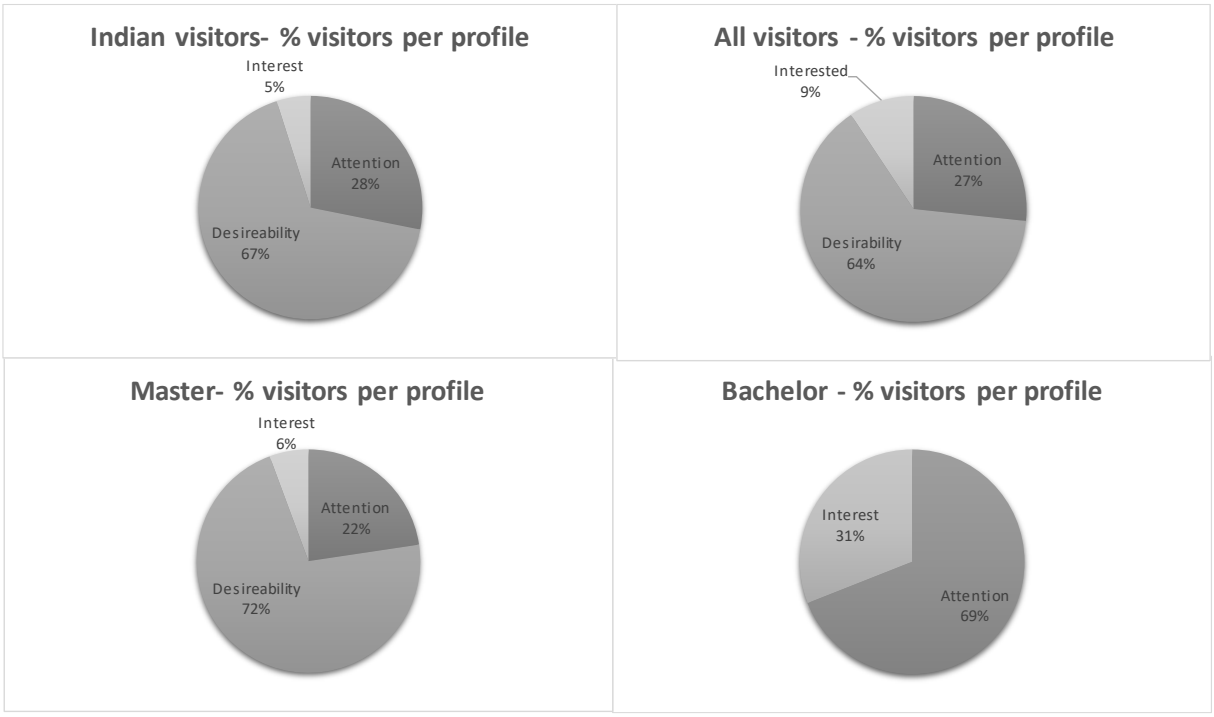

Figure 7 – visitors' disribution in behaviourl profiles for all analyses

Each graph in figure 8 demonstrates visitor distribution to the three behavioural profiles, namely *Attention*, *Desirability* and *Interest* profiles. The three discovered behavioural profiles occur consistently among all visitors, Indian visitors and Master study level. However, a

behaviour profile, namely *Desirability*, is absent among visitors interested in Bachelor studies. The reason for this is described in 4.3.3 section of this report, which is related to absence of Eligibility check for bachelor study level. Moreover, percent of visitor distribution to each discovered behavioural profile remains approximately same when looking at the profiling of all visitors, Master and Indian visitors. This goes to show that even though the actual number of visitors varies from one analysis to another but the proportion of visitor distribution remains approximately the same. In other words, Desirability profiles are consistently the largest profiles followed by Attention profiles and then Interest profiles. This inference is consistent when looking at analysis outcome of all visitors, Indian visitors and visitors interested in Master studies.

The Behavioural profiles discovered among Indian visitors looks fairly similar to the behavioural profiles among all visitors and visitors interested in master studies. At first glance, this seems to be self-explanatory but considering the fact that in analysing Indian visitors, no distinction made between two study levels. Thus, when evaluating the distribution of Indian visitors to each study level, it becomes clear that majority of Indian visitors are interested in Master studies. To be exact 96% of Indian visitors are interested in master studies and the remaining 4% are interested in Bachelor studies. The distribution of visitors interested in two study levels can be seen in figure 9. It can be assumed that the three discovered behavioural

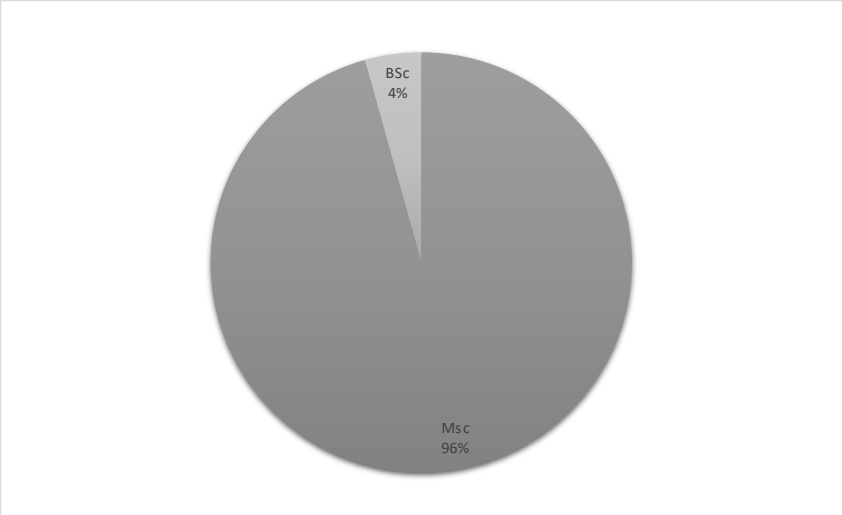

Figure 8 – Percent of Indian visitors interested in Master vs Bachelor studies

profiles among Indian visitors are largely influenced by the large proportion of Indian visitors who are interested in master studies. Thus, providing evidence as to why Behavioural profiles of Indians bears high resembles to the behavioural profiles of all visitors and visitors interested in Master studies. As a result, it can be hypothesized that Behavioural profiles are the same across countries and cultures or at least the behavioural profile of Indian users are no different than all visitors. Therefore, Behavioural profiles are most likely dependant on study levels rather than the country factor.

The outcome of analyses suggests that few behaviours are dominant and their pattern is quite striking in each Behavioural profile. These dominant Behavioural features are Eligibility check, Brochure Request and Question via web form. These behaviours are striking behaviours that became the basis for naming and describing all discovered Behavioural profiles in this paper. The frequency and pattern of aforementioned behaviours are so bold among all profiles that they potentially could be used for allocation of new visitors to one of three

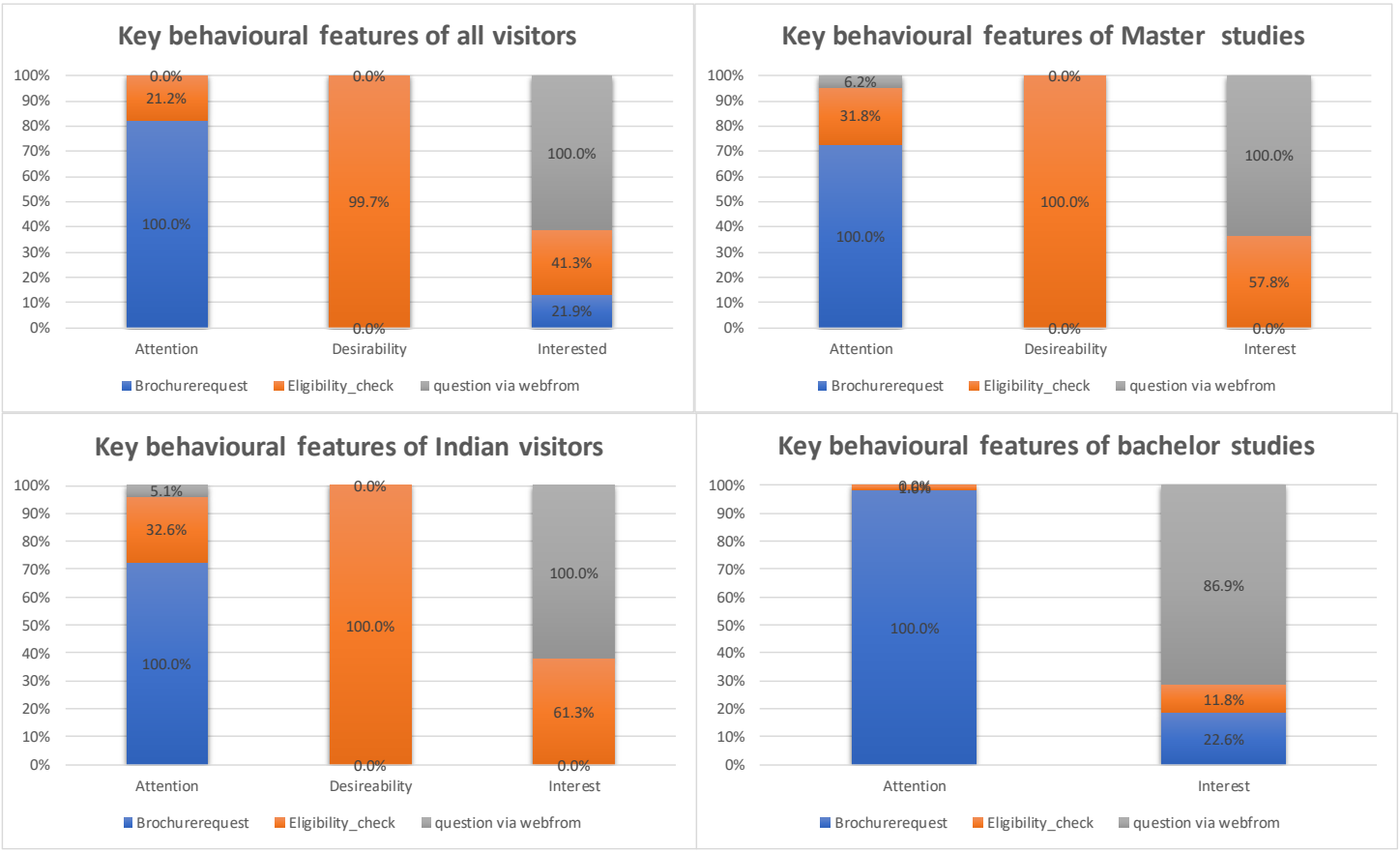

Figure 9 – Distribution of Key behavioural features in each profile across analyses

behavioural profiles discovered. Thus, these behavioural features can be called Key behavioural features. Graphs of figure 10 illustrate the distribution of Key behavioural features in each profile across analyses. Each colour represents a Key behaviour, namely colour blue represent Brochure request and colour grey represent Question via web form and colour orange represent Eligibility check behaviour. Each stacked bar in each of graph represents a profile. Simply by looking at the distribution of Key behavioural features in each profile, one can observe a certain pattern in terms of Key behavioural among all profiles across analysis. This holds true for all analyses except visitors interested in Bachelor studies. As mentioned in the previous chapter, the reason for this is absence of Eligibility check for visitors interested in bachelor studies. Yet, one can observe despite that the distribution of Key behavioural is approximately the same among visitors interested bachelor studies as for instance visitors interested in Master studies. For example, in the Attention profile among visitors interested in both master and bachelor, 100% of visitors manifest a Key behaviour feature, namely Brochure request. In addition to *Key behavioural* features, there are eight other behavioural features that website visitors manifest although infrequently. Such behavioural features do not follow any particular or consistent pattern or ratio across analyses. Thus, patterns or ratio of such behavioural features in each analysis is unique to the same analysis. These behavioural features can be called *Micro Behavioural* features. Example of such behavioural features are *Frequently asked questions, PDF download, Managed CTA Click* and *Managed CTA Display.* For instance, the frequency ratio of Micro behavioural features is different between Attention profile among all visitors and Indian visitors. Thus, it could be assumed that manifestation of each *Micro Behaviour* feature is dependent on visitors' interested study level and their country of origin. This is an interesting insight as it can be used to improve the effectiveness of marketing campaigns where a specific advertisement can be shown to visitors depending on their country and their interested study

levels. The distribution of Micro behavioural features among three behavioural profiles across analyses can be seen in figure 11 where they are compared in a side-by-side manner.

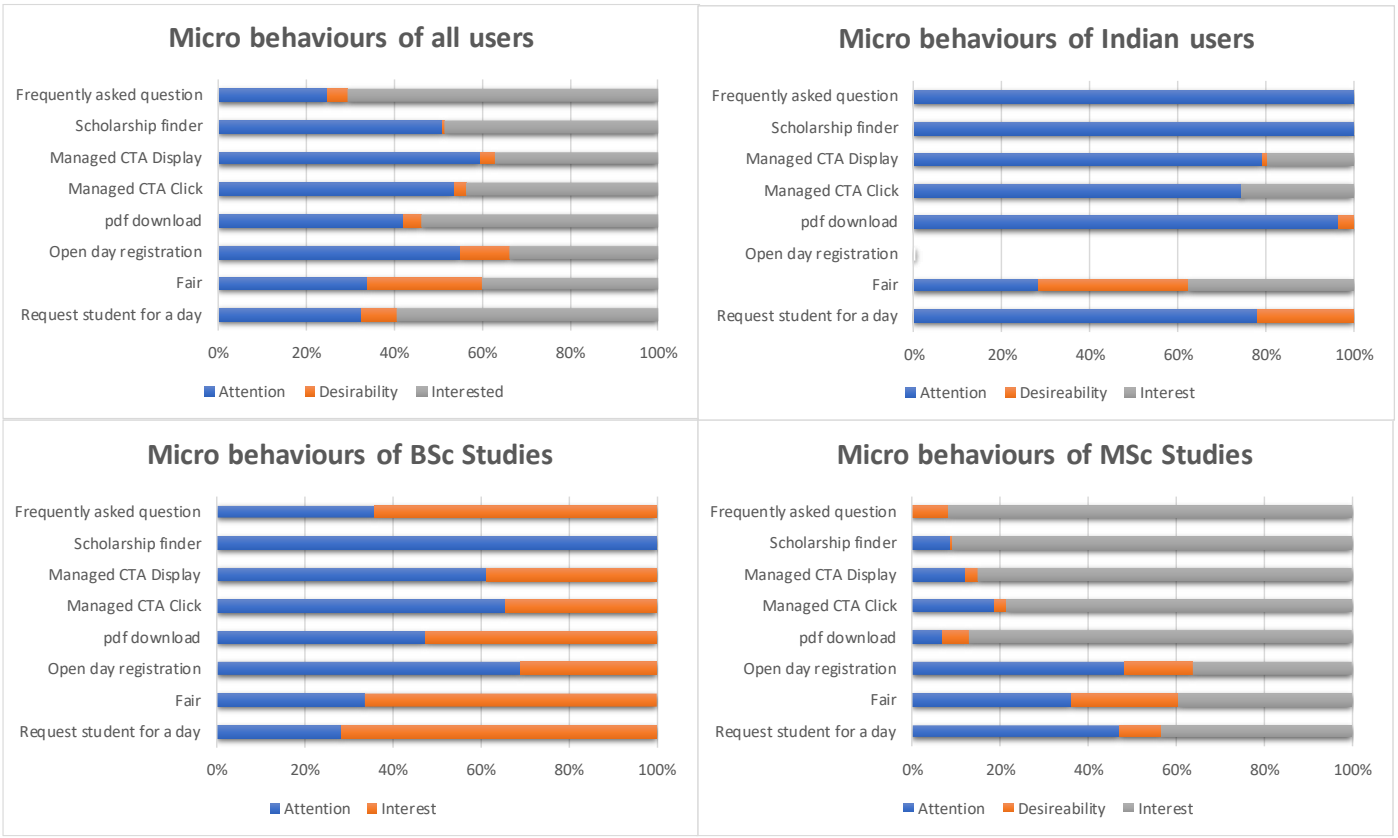

Figure 10 - Side by side comparison of Micro behavioural features among behavioural profiles across all analyses

In terms of Behavioural source features, the available data is simply not enough to make any conclusion that is robust or valid about entry points of visitors. The Behavioural source variance among profiles and across analyses, as shown in figure 12, is very small. And yet, if such small variance is taken as an indication of small yet true behavioural source of website visitors of the University of Twente, then the following conclusions can be made.

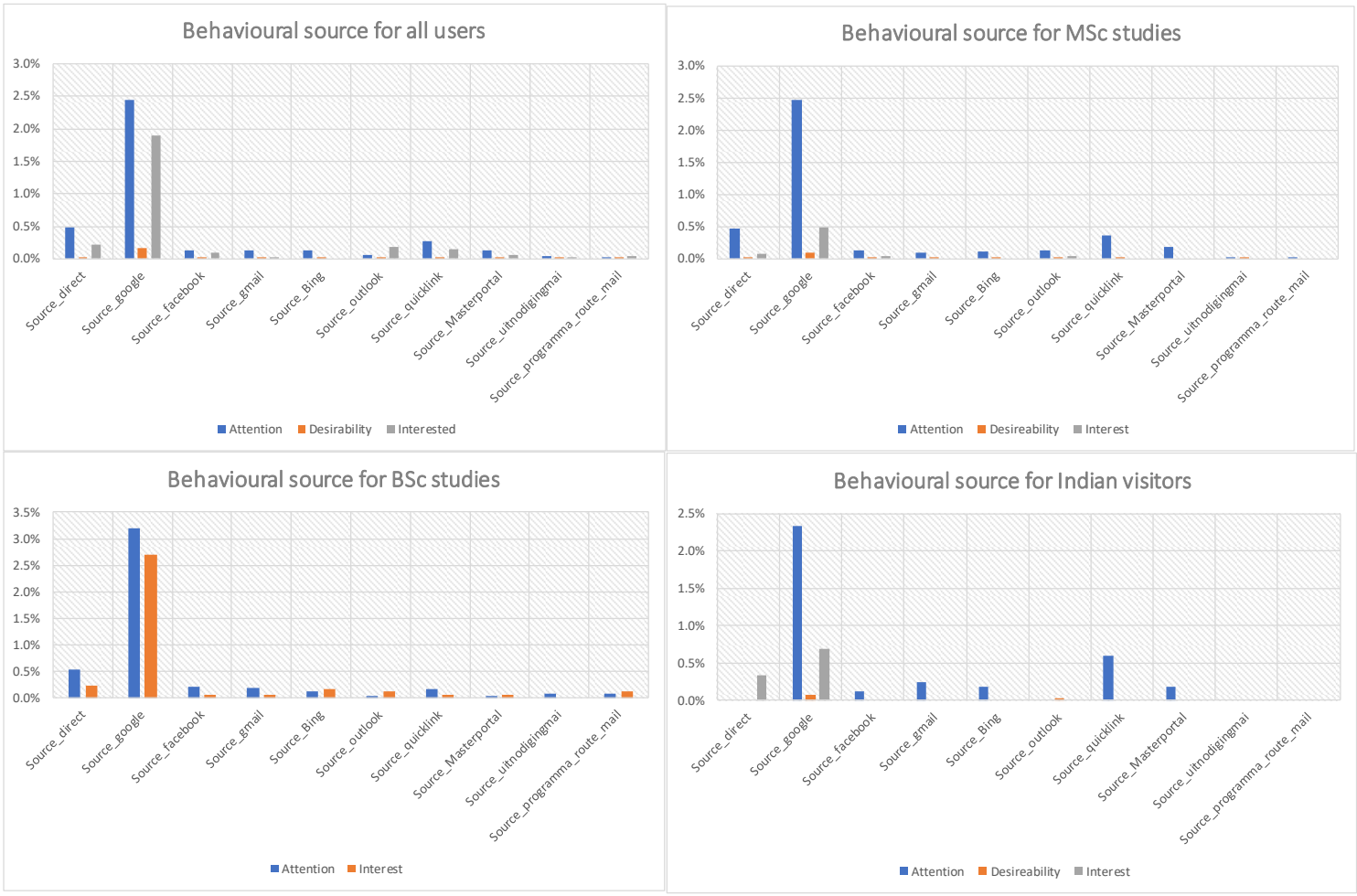

Figure 11 – Distribution of *Behavioural source* among all profiles across analyses

The analyses revealed a repetitive pattern in terms of Behavioural source feature among visitors of Attention profile. It suggests that majority of members in that profile enter UT's website from Google. This outcome is consistent across all analyses. Moreover, there is almost no information about how members of Desirability profile enter UT's website. However, the same cannot be said about visitors interested in Bachelor studies. A small portion of such visitors, who belong to the Desirability profile, enter UT's website via Google. A reasonable explanation would be that since Eligibility check is not available in pages of Bachelor studies, then visitors who would like to do the test,

find their way by searching on Google. Overall, more research is required to reach robust conclusions regarding entry points of UT's website visitors. Perhaps dataset used in this paper is incomplete, for instance, there might be a problem in data collection or simply this paper did not take other entry points into account that visitors use to get to their desired pages.

4.4. Clustering Validation

Previous sections of this chapter focused on discovering behavioural profiles of website visitors at the University of Twente. All profiles were described in terms of behavioural features, which were defined in chapter 3, and the consistency of profiles were evaluated by controlling for country and study level factors. Although the discovered behavioural profiles proved to be consistent across analyses, the validity of discovered profiles were not assessed. This section describes the outcome of clustering validation outlined in chapter 3. Validation of discovered behavioural profiles is just as important as discovering and describing them. As a result of validating Behavioural profiles, conclusions based on profiles becomes more robust and valid. In total three different validity test are performed, namely Kruskal-Wallis, Silhouette score, and Cross-validation. Here below, the outcome of each test is described in the same order that they are mentioned.

4.4.1. Kruskal-Wallis test

This test evaluates the validity of profiles by statistically prove that each profile is different than another one. This test indicates that the three discovered behavioural profiles are different at least in one area from one another. The distribution of behavioural features for each analysis does not follow the normative curve and cannot be transformed to a parametric one, thus it is non-parametric data. Therefore, ANOVA is not suitable but Kruskal-Wallis test can be used for such data. The outcome of Kruskal-Wallis test on profiles of each analysis can be seen in table 20. Overall, results indicate that indeed all profiles of each analysis are statistically

different in at least one behavioural feature. The alpha level for this analysis is set at 5% (Alpha level = 0.05).

Table 16

Kruskal-wallis test of each clustering solution

all visitors		Bachelor		Master		Indian	
Statistic	P-value	Statistic	P-value	Statistic	P-value	Statistic	P-value
9658.3701	0	3275.83	0	5643.47	0	9876.255	0

4.4.2. Silhouette score – Cluster Homogeneity

In this section, cluster homogeneity of profiles discovered in each analysis is evaluated by means of silhouette score. Such cluster homogeneity evaluates the validity of profiles by looking at how closely each visitor is located to its profile centroid. The silhouette score for each analysis is shown in table 17. As explained in chapter 3 section 3.1, a silhouette scores close to 1 means profiles are quite valid and as result, cases (who are visitors in this paper) are quite cohesive in each profile. When the score is closer to -1, then the profiles are not as valid or robust.

Table 17

Silhoutte score of analyses

all visitors	Bachelor	Master	Indian
.907477375569034	.721476730406697	.932456403992194	.921342756761905

The silhouette score of all visitors, master and Indian visitors are all above 0.9, which is indicative of a very strong result. The silhouette score for bachelor is 0.72, although not quite as high as the other analyses but it is still indicative of a good solution. In general, any score above 0.5 is indicative of good clustering solution as it is closer to 1 than 0 or -1, and the silhouette score for all analyses in this paper are above the 0.5. Overall, silhouette score indicates that the behavioural profiles are valid and robust.

4.4.3. Cross-Validation

In this section, cross-validation for each analysis is performed by using the hold-out method described in chapter 3 section 3.3. This variation of Machine Leaning validation evaluates the appropriateness of the chosen number of profiles for analyses. This is done by randomly dividing the original data into two sub-samples with ratios of 75% (test dataset) to 25% (Train dataset). By recalculating the Wcss score for each sub-sample, one can find out the appropriate number of profiles for each analysis. The figure 13and 14, visualizes Wcss score of each sub-sample of all visitors.

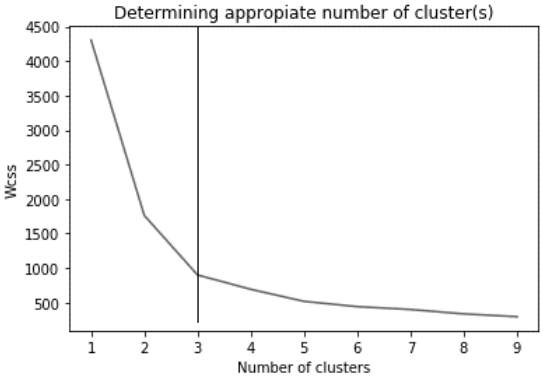

Figure 13 – Wcss score of test dataset

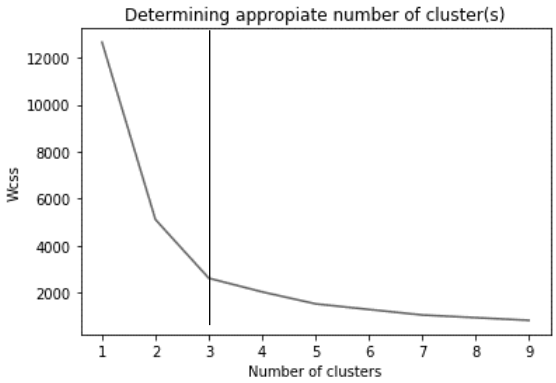

Figure 12 – Wcss score of Train dataset

Each of these graphs shows that the appropriate number of profiles (indicated by a vertical line) remains the same. The graphs indicate that Wcss score of each sub-sample does not change dramatically beyond the points indicated by vertical lines. Moreover, the same Cross-validation method performed on other analyses, namely Indian visitors, visitors interested in bachelor and master studies, suggest that the number of profiles remains the same as the ones calculated in section 4.2 of paper. For instance, the cross-validation on the visitors interested in bachelor studies yields 2 clusters as the appropriate number of clusters.

Furthermore, the recalculated the silhouette scores for the clustering solution of the two sub-samples of all visitors' datasets remains approximately the same. The silhouette score for the Test dataset is 0.8932 and for the Train dataset is 0.91266. Both scores indicate a strong

solution and thus it can be said that the behavioural profiles of all visitors are quite valid and robust. The same is true when the silhouette score is recalculated for other segments.

5. Discussion

This paper goal of this paper was to provide an overview of approaches to customer profiling and customer attribute prediction. This was achieved by reviewing the various literature regarding Machine Learning and customer attributes within the context of marketing. The outcome was derived in form of a framework that can be found in chapter 2 section 10. Furthermore, a model was developed to account for nature of data and incorporating Multi-step segmentation approach to user profiling. It is derived by combining *Hybrid* user profiling type proposed by Khosrow-pour (2009) and *Hybrid* segmentation approach proposed by Dolnicar (2008) . This model can be found in chapter 2 section 10. Moreover, the goal in the paper was to discover behavioural profiles of website visitors in higher education, which in this paper was the University of Twente. Using the proposed framework and model in this paper, three behavioural profiles were discovered, namely *Attention*, *Interest* and *Desirability* profile. All of these profiles are distinguishable by behavioural features that were called *Key behavioural* features, namely *Eligibility check*, *Question via web form* and *Brochure download.* The pattern and manifestation frequency of the *Key behavioural* features are unique for each of the aforementioned profile and consistent across all analyses. However, the same could not be said for other behavioural features. The remaining behavioural features were manifested infrequently and their patterns were not consistent across analyses or profiles. Such behavioural features were named *Micro Behavioural* features in this paper.

Typically, Machine Learning has been mainly used in the field of IT. However, due to its popularity ML found its way in Business field. For instance, (Yao et al., 2010) used Machine Learning to increase spending amount of their existing customers. They achieved this by identifying their customer preferences, such as purchasing behaviour and spending amount to

define user profiles. In contrast, this paper utilizes ML to find and understand what visitors' behaviours (Explicit or Implicit in nature) translate to a desired conversion within the marketing context.

Moreover, prior research used viewing pattern of website visitors as features (such as time spend on pages, click pattern) to create user profiles but this paper uses the actual interaction of visitors with the website (defined as behaviour in this paper) to discover user profiles. This was achieved by using multiple data sources, namely Web data and Business data. Feature selection is quite important in Machine Learning for generating meaningful insights. Often viewing patterns of website visitors are distorted and does not represent the true intention and interest of visitors. They often end up in wrong pages and thus they have to go back and forth until they find what they seek. However, behavioural features used in this research provide a better indication of visitors' intentions and interests as it is less distorted by noises that exist in visitors' viewing pattern.

Moreover, analysing all website visitors regardless of their country or study levels seem to generate distorted behavioural profiles. The analysis on all website visitors suggest three behavioural profiles, namely *Attention*, *Interest* and *Desirability* exists. However, when behavioural profiles are created first Apriori then Posteriori, segments such as visitors interested in Bachelor studies, the result is different. It seems that visitors interested in bachelor studies are missing the Desirability profile. This provides evidence that by applying the *Multi-step user profiling process* model proposed in chapter 2, profound insight can be gained from data that was not possible to know by using approaches used in the literature.

Furthermore, utilizing the *Multi-step user profiling process* model led to the discovery of two categories of behaviours that would have been overlooked if analyses were done using a one-step approach to create user profiles. For instance, the outcome of analyses suggest that certain behaviours are manifested frequently, have patterns that is repeated across analyses .

Such behaviours are called *Key Behaviours* in this paper. The patterns of such behaviours do not vary significantly when factors such as country or study levels are controlled. However, the patterns of another group of behaviours, namely *Micro behaviours*, changes dramatically when factors such as country and study levels are controlled. This insight indicates that identification and distinguishing the two aforementioned behaviours are important as using only one to create user profiles could lead to distorted and generic conclusions.

6. Conclusion

This paper set to find out various approaches and methods that allows customer segmentation based on customer attributes data using Machine Learning. By reviewing various literature regarding customer attributes and Machine Learning, this paper developed a framework that provides an overview on customer segmentation and customer attribute prediction. Additionally, this paper proposed a model that describe various processes to customer profiling based on different user profiling types and segmentation approach. The model is called *Multi-step user profiling process* model. Utilizing the model for creating customer profiles can generate profound result comparing to one-step approaches used in the literature, as this model takes the nature of customer data (Implicit vs Explicit) into account and make use of two-step segmentation approach (known as Hybrid, which is various combination of Apriori and Posteriori approach). Furthermore, the goal in the paper was achieved by utilizing the proposed framework and model to create behavioural profiles of website visitors for the University of Twente. The behavioural profiles discovered as a result of utilizing the proposed framework and model in this paper, are significantly different than the behavioural profiles discovered when the approaches used in the literature are utilized. Such differences are visible on certain behavioural features, that are named *Micro behaviours* in this paper. In total, three behavioural profiles were discovered, namely *Attention*, *Interest* and *Desirability* profile. The findings of this paper have several insightful theoretical and practical implications and yet

at the same time, they are subjected to several limitations, which can be used as an indication for future research.

6.1. Theoretical Implication

This paper lay the foundation for user profiling and customer attribute prediction using Machine Learning. The Framework proposed in this paper provides an overview of approaches using various combination of Machine Learning techniques and customer attributes. In addition, this framework provides guidance to researchers in Business field as to when, where and to a certain degree how each of the two main categories of Machine Learning could be used. This is especially true within Marketing area of Business field.

Furthermore, this paper proposes a model, called the *Multi-step user profiling process* model, that could be used for user profiling. Researchers can gain profound insight as a result of using the proposed model. This model is sensitive to small variation among different users as a result of segregating them based on user profiling types (explicit data vs implicit data) and using the Two-step segmentation approach proposed by Dolnicar (2008). As a result of such combination, this model does not have the shortcomings of one-step segmentation approach nor each one of user profiling type alone. Thus, it provides a profound insight and yet it is easy to interpret. In addition, the user profiles are not trivial but at the same time, they are not too complex that is beyond comprehension.

In conclusion, the framework and model proposed in this paper are complimentary for the purpose of user profiling. By using them in combination, a researcher could gain valuable and profound insight. Furthermore, the proposed framework provides an overview of various approaches to customer attribute prediction. Researchers can use this overview to categorise previous research on customer attribute prediction in order to find the gap for future research. By utilizing aforementioned framework and model, this paper was able to come to valuable

conclusions that could be used to improve marketing practices and they are described in the following section.

6.2. Practical Implication

The outcome of this paper suggests the existence of two group of behavioural features. They are *Micro behaviours* and *Key behaviours*. Understanding and realizing the two categorize are important as each user profile manifest a different pattern in terms of Micro behaviours in this paper. However, user profiles are not so much different in terms of Key behaviours. In addition, Micro behavioural patterns of user profiles vary across different segments. Thus, this means user profiles across segments are different in terms of Micro behaviours but not so much in terms of Key behaviours. As a result, Marketing campaigns should be designed in a way that corresponds to the micro behaviour of user profiles across different segments rather than focusing on commonalities (key behavioural pattern across segments). By doing so, the effectiveness and efficiency of marketing campaigns can be improved.

Furthermore, by utilizing the framework and model proposed in this paper, higher educations can identify behavioural profiles and the specific behavioural patterns of visitors that lead to a desired conversion. By identifying such behavioural patterns for each behavioural profile, higher educations can modify re-targeting campaigns to improve effectiveness and efficiency of their advertisements. Furthermore, insight on behaviours leading to a desired conversation, provide valuable insight on user profiles. Such insight can be used to figure out the needs and desires of each profile. By identifying the needs and desires of each profile, separate marketing campaigns can be created tailored specific to members of each user profile. In addition, insight on existing visitors can be used to realize characteristics of visitors who are attracted to the higher education. By realizing such characteristics, Lookalike marketing campaigns can be created to drive more relevant visitors to the higher education website.

Moreover, the outcome of this paper based on the data of the University of Twente suggests that the behavioural profiles of website visitors are country independent, or at least in case of India. This goes to show that marketing campaigns based on trivial visitor segmentation, such as geographic profiling one-step segmentation, might not be ideal. This point is supported by the fact that behavioural profiles of Indian visitors are no different than visitors interested in master studies. Thus, potentially indicating that behaviours of website visitors are not dependent on countries but rather on the interested study levels (master vs bachelor).

6.3. Future research and Research Limitation

This paper laid the foundation for future research, on more narrow and specific goals for user profiling. Future research could find out the sequence of manifested behaviours among behavioural profiles or possibly create a prediction model that could classify new visitors to one of three discovered behavioural profiles. The framework and model proposed in this paper, provide guidance on how user profiles can be created and customer attributes prediction. Thus, providing guidance for developing predictive models for a certain behavioural trait or as mentioned before, classification of new visitors to one of the profiles.

Moreover, future research can investigate *Micro behavioural* features of visitors of multiple countries who are interested in same study level, to see the variation of *Micro behaviours* among visitors of each country. For instance, conducting a research on a *Micro behaviour* feature such as *FAQ* to see if certain frequently asked questions are more interesting to visitors of a particular country. Such information can be used to enrich advertisement message in marketing campaigns. Furthermore, future research can evaluate user profiles to find out how behavioural profiles are different from one another, or in other words, what behavioural features are statistically different from the other profiles. By realizing such information, researchers can realize the most important behavioural features of each profile and in turn, use them for marketing campaigns.

Future research can find the optimal level homogeneity of "between members" in clustering for Behavioural targeting in high dimensional and high-volume data. Furthermore, future research can be conducted on multiple countries to see if behavioural profiles of website visitors of higher education are the same or indeed behavioural profiles of countries varies.

In-depth research based on results and conclusions of this paper could provide a more accurate picture of visitor's behaviour as well as their granular behavioural patterns. Therefore, research using techniques such as path analysis, association rule, network analysis and time sequence analysis could potentially unravel interesting insight. Moreover, future research can investigate how website visitors find their way to their desired pages. Insight generated as result of such research could shed light on the Behavioural source of profiles, which could be taken as used for advertisement placement.

Future research can be conducted to create prediction model, one that is able to classify new visitors to one of three behavioural profiles mentioned in this paper. Such model can be a valuable tool to marketing department, as they can use it to deliver appropriate information to new visitors based on their behavioural profiles.

The conclusion of this paper is limited by veracity and variety of datasets used. In addition, the quality level of the data used in this paper is not clear, therefore a confirmatory research, using the same methods based on the website visitors of UT in a 2-3 years' time could potentially overcome this issue. Moreover, the conclusion of the paper could be improved if each discovered behavioural profile is evaluated in terms of numbers of visitors who were accepted as students rather than visitors who took Eligibility check. Such insight can help to find out the most valuable profile to focus marketing activities.

The conclusions of this paper are not generalizable. However, the proposed framework and model could be used with a different dataset. In addition, confirmatory research using the same approach in this paper on a new dataset could provide evidence about the degree of

generalizability of conclusions made in this paper. Moreover, the features available in the dataset determines the scope, precision of conclusion, therefore this study is limited by the richness of the raw data sources. More study on features and inclusion or exclusion of features from different data sources could potentially reveal different and interesting results.

7. Reference

Ahmed, A., Low, Y., Aly, M., Josifovski, V., & Smola, A. J. (2011). Scalable distributed inference of dynamic user interests for behavioral targeting. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, 114. https://doi.org/10.1145/2020408.2020433

Araniti, G., De Meo, P., Iera, A., & Ursino, D. (2003). Adaptively Controlling the QoS of Multimedia Wireless Applications Through "User Profiling" Techniques. *IEEE Journal on Selected Areas in Communications*, *21*(10), 1546–1556. https://doi.org/10.1109/JSAC.2003.815226

Araya, S., Silva, M., & Weber, R. (2004). A methodology for web usage mining and its application to target group identification. *Fuzzy Sets and Systems*, *148*(1), 139–152. https://doi.org/10.1016/j.fss.2004.03.011

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. https://doi.org/10.1016/J.PATCOG.2012.07.021

Asanka, P. (2017). Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach. Retrieved from https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera/

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

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

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. https://doi.org/10.1016/j.knosys.2016.05.002

Bureau, I. advertising. (2014). A guide to Online behvaiour advertising.

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, L., & Yu, P. S. (2012). Behavior computing: Modeling, analysis, mining and decision. *Behavior Computing: Modeling, Analysis, Mining and Decision*, 1–374. https://doi.org/10.1007/978-1-4471-2969-1

Castelluccia, C. (2012). European Data Protection: In Good Health?, 21–34. https://doi.org/10.1007/978-94-007-2903-2

Chen, J., & Stallaert, J. (2014). an Economic Analysis of Online Advertising Using Behavioral Targeting. *MIS Quarterly*, *38*(2), 429-A7. https://doi.org/10.2139/ssrn.1787608

Chopra, P. (2012). Behavioral Targeting: the most underused technique in today's marketing. Retrieved from https://vwo.com/blog/behavioral-targeting/

Chorianopoulos, A. (2016). *Effective CRM ussing predictive analysis*. Wiley.

Click-through rate (CTR): Definition. (2017). Retrieved from https://support.google.com/adwords/answer/2615875?hl=en

CLUTO - Software for Clustering High-Dimensional Datasets. (2006). Retrieved from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

Conover, W. J., & Iman, R. L. (1979). On Multiple-Comparisons Procedures. *Technical Report LA-7677-MS*, 1–14. Retrieved from

http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-07677-MS

Conversion Rate. (2017). Retrieved from http://www.marketingterms.com/dictionary/conversion_rate/

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric Statistics for Non-Statisticians*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118165881

Cufoglu, A. (2014). User Profiling-A Short Review. *International Journal of Computer Applications*, *108*(3), 9. Retrieved from http://research.ijcaonline.org/volume108/number3/pxc3900179.pdf

Dao, T. B. H., Duong, K. C., & Vrain, C. (2015). Constrained minimum sum of squares clustering by constraint programming. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9255*, 557–573. https://doi.org/10.1007/978-3-319-23219-5_39

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. https://doi.org/10.1016/j.ins.2015.06.039

Deane, J. K., Meuer, T., & Teets, J. M. (2011). A longitudinal analysis of web surf history to maximise the effectiveness of behavioural targeting techniques. *International Journal of Electronic Marketing and Retailing*, *4*(2–3), 117–128. https://doi.org/10.1504/IJEMR.2011.043037

Dibb, S., & Simkin, L. (1996). *The market segmentation workbook : target marketing for marketing managers*. London: Routledge.

Directive Group. (2017). Psychographic market segmentation. Retrieved from https://www.localdirective.com/what-we-do/market-segmentation/psychographic/

Doig, C. (2015). Topic Modeling with Python. Retrieved from

https://www.youtube.com/watch?v=BuMu-bdoVrU&list=RDQMlyJh3a_LBaU

Dolnicar, S. (2008). Market segmentation in tourism. *Tourism Management: Analysis, Behaviour and Strategy*, 129–150. https://doi.org/10.1079/9781845933234.0129

Dolničar, S. (2004). Beyond "Commonsense Segmentation": A Systematics of Segmentation Approaches in Tourism. *Journal of Travel Research*, *42*(3), 244–250. https://doi.org/10.1177/0047287503258830

Dunn, O. J. (1964). Multiple Comparisons Using Rank Sums American Society for Quality Stable URL : http://www.jstor.org/stable/1266041 Linked references are available on JSTOR for this article : *Technometrics*, *6*(3), 241–252.

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

Fennemore, P. (2011). Market segmentation : how does it work with online social networks ?

FESBAL. (2013). What are cookies? Retrieved from http://www.bancodealimentos.es/bancos/aprende/queson

GODOY, D., & AMANDI, A. (2005). User profiling in personal information agents: a survey. *The Knowledge Engineering Review*, *20*(4), 329. https://doi.org/10.1017/S0269888906000397

Goldfarb, A., & Tucker, C. E. (2011). Privacy Regulation and Online Advertising. *Management Science*, *57*(1), 57–71. https://doi.org/10.1287/mnsc.1100.1246

Gong, X., Guo, X., Zhang, R., He, X., & Zhou, A. (2013). Search behavior based latent semantic user segmentation for advertising targeting. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 211–220. https://doi.org/10.1109/ICDM.2013.62

Hand, D. . (1981). *Discrimination and Classification*. Wiley.

Hui, L. (2017). Which machine learning algorithm should I use? Retrieved from https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

Internet Users. (2017). Retrieved from http://www.internetlivestats.com/internet-users/

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

Jaworska, J., & Sydow, M. (2008). Behavioral Targeting in On-Line Advertising: An Empirical Study, 62–76.

Kanoje, S., Girase, S., & Mukhopadhyay, D. (2014). User Profiling Trends, Techniques and Applications. *International Journal of Advance Foundation and Research in Computer*, *1*(11), 2348–4853.

Khosrow-Pour, M. (2009). *Encyclopedia of information science and technology*. Hershey, PA Idea Group. Retrieved from http://books.google.com/books?id=J0DkQwAACAAJ&printsec=frontcover%5Cnpapers://5e1137fc-0038-41d8-bf7a-dd54318682de/Paper/p146

Kim, L. (2017). Click-Through Rate (CTR): Understanding Click-Through Rate for PPC.

Kuflik, T., & Shoval, P. (2000). Generation of User Profiles for In formati on Filtering-Research Agend. *SIGIR Conference on Research and Development in Information Retrieval*, 313–315. https://doi.org/10.1145/345508.345615

Leon, G. (2016). Daypart Trends and Best Practices for DRTV Campaigns. Retrieved from http://www.hawthornedirect.com/blog/daypart-trends-and-best-practices-for-drtv-campaigns/

Lewis, E. S. E. (1908). The History of Advertising. *Financial Advertising*.

Local Drirective. (2017). Behavioural Segmentation of Your Targeted Market. Retrieved from https://www.localdirective.com/what-we-do/market-segmentation/behavioral/

Lu, X., Zhao, X., & Xue, L. (2016). Is Combining Contextual and Behavioral Targeting Strategies Effective in Online Advertising? *ACM Transactions on Management Information Systems*, *7*(1), 1–20. https://doi.org/10.1145/2883816

Morning, E., & Morning, E. (2017). KANTAR Daypart definitions :, 59.

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

Pariser, E. (2011). *The Filter bubble*.

Plummer, J., Rappaport, S., Hall, T., & Barocci, R. (2007). *The Online Advertising Playbook: Proven Strategies and Tested Tactics from the Advertising Research Foundation*. Retrieved from http://www.rmit.eblib.com.au/EBLWeb/patron?target=patron&extendedid=P_309746_0&

Pollard, K., & van der Laan, M. (2002). A method to identify significant clusters in gene expression data. *Proceedings of SCI, II*, 318–325. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.6332&rep=rep1&type=pdf

Poo, D., Chng, B., & Goh, J.-M. (2003). A Hybrid Approach for User Profiling. *SciencesNew York*, *4*(C), 103–111. https://doi.org/10.1109/HICSS.2003.1174242

Rawal, P. (2013). AIDA Marketing Communication Model: Stimulating a purchase decision in the minds of the consumers through a linear progression of steps. *International Journal of Multidisciplinary Research in Social & Management Sciences*, (1), 37–44.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of

cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schneider, J. (1997). Cross-validation. Retrieved from https://www.cs.cmu.edu/~schneide/tut5/node42.html

Solutions, S. (2017). Kruskal-Wallis Test. Retrieved from http://www.statisticssolutions.com/kruskal-wallis-test/

Terradata. (2015). Progressing Toward True Individualization.

Tsai, C. Y., & Chiu, C. C. (2004). A purchase-based market segmentation methodology. *Expert Systems with Applications*, *27*(2), 265–276. https://doi.org/10.1016/j.eswa.2004.02.005

Tu, S., & Lu, C. (2010). Topic-based user segmentation for online advertising with latent dirichlet allocation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6441 LNAI*(PART 2), 259–269. https://doi.org/10.1007/978-3-642-17313-4_26

Wijaya, B. S. (2012). The Development of Hierarchy of Effects Model in Advertising. *International Research Journal of Business Studies*, *5*(1), 73–85. https://doi.org/10.1555/IRJBS.V5I1.98

Williams, R. (2004). Multiple / Post Hoc Group Comparisons in ANOVA. *Sociology Graduate Statistics I, University of Notre Dame*, 1–5. Retrieved from http://nd.edu/~rwilliam/stats1/x53.pdf%5Cnhttp://nd.edu/~rwilliam/stats1/

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. https://doi.org/10.1145/1592748.1592751

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How much can behavioral targeting help online advertising? *Proceedings of the 18th International Conference on World Wide Web - WWW '09*, 261. https://doi.org/10.1145/1526709.1526745

Yao, Z., Eklund, T., & Back, B. (2010). Using SOM-Ward Clustering and predictive analytics for conducting customer segmentation. In *Proceedings - IEEE International Conference on Data Mining, ICDM* (pp. 639–646). https://doi.org/10.1109/ICDMW.2010.121

Zhou, Y., & Mobasher, B. (2006). Web user segmentation based on a mixture of factor analyzers. *E-Commerce and Web Technologies*, 11–20. Retrieved from http://link.springer.com/chapter/10.1007/11823865_2

8. Appendixes

Appendix 1

Table 18 Summary of User profile types by Khosrow-pour (2009)

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 2

DYNAMIC	***Dynamic Content Profiling*** refers to gathering of information based on the dynamic changes in the behaviour of the user and filtering only those that represent the user's profile.	***Dynamic Collaborative Profiling*** refers to organising users with similar behaviour into peer groups based on the user's profile and filtering information pertaining to group's interest.
STATIC	***Static Content Profiling*** refers to the gathering of static information regarding the user only.	***Static Collaborative Profiling*** refers to explicitly organising users with similar behaviour into peer groups through user explicit request.
	CONTENT	**COLLABORATIVE**

Figure 14- Variation of Hybrid user profiling methods by Poo et al. (2003)

Appendix 3

Table 19 Summary of user profile methods by Cufoglu (2014)

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

			Explicitness		Implicit				
Machine learning algorithm (ML)	Data type condition	ML Technique	Demographic	Geographic	Daypart	Affinity	Purchased-based	Behavioural	Psychographic
Unsupervised	Low volume & low dimensional (<1000 entries)	***Hierarchical Clustering + Non-hierarchical Clustering***	Segment users based on Demographics	Segment users based on their geographical location	Segmenting users based on time consumption/Usage	Segmenting users based on similarity in brand affinity	Segmenting users based on their purchasing behaviour	Segmenting users based on user behaviour	Segment users based on their psyche attributes
	High volume & high dimensional (+1000 entries)	***Wcss + Non-hierarchical Clustering***							
Supervised	Categorical dependent feature	***Classification***	Predicting demographic of users	Predicting location of users	Predicting user's time of consumption/Usage of a product or service	Predict user's interest in a brand	Predicting user's probability of purchasing a product/service	Predicting user behaviour profile/a desired behaviour	Predicting user psyche
	Continuous dependent feature	***Regression***							

Figure 15 – Framework to user profiling and customer attribute prediction using ML