UNIVERSITY OF TWENTE.

Faculty of Behavioural, Management, and Social Sciences

BEHAVIOURAL PROFILES OF POTENTIAL STUDENTS AS BASIS FOR MORE EFFECTIVE UNIVERSITY RECRUITING

F.J. Kuiper

Master's Thesis
MSc. Business Administration - Strategic Marketing and Business Information
December 2018

Supervisors:
Dr. Efthymios Constantinides
Dr. Sjoerd de Vries

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

Management Summary

Purpose - Large amounts of data are being collected at a dramatic pace. However, organizations often have difficulties to extract knowledge from data and selecting appropriate Machine Learning and User Profiling approaches to fully harness the potential of Behavioural Targeting techniques. Moreover, (university) marketing departments often lack a fundamental understanding on data-driven segmentation methodologies. In addition, lack of research and cases make it difficult to develop profiles of potential students based on their search behaviour and other characteristics. This paper aims to develop a framework of Unsupervised Machine Learning (UML) algorithms for User Profiling with respect to important data properties. Moreover, the aim is to discover high converting behavioural profiles among Dutch website visitors of the University of Twente (UT) interested in UT Master studies.

Methodology - A literature review is conducted and the process of Knowledge Discovery in Databases is used as a research methodology. Data was collected between October 2016 and August 2017 from the UT CRM-system and Google Analytics. Complete Linkage and K-modes are used for data analysis in combination with Hybrid User Profiling.

Findings - A framework is proposed of UML algorithms for User Profiling. It provides two-stage clustering approaches for categorical, numerical, and mixed types of data with respect to the data size and data dimensionality. Six behavioural profiles were discovered of which two are most significant in terms of conversions. In addition, a model is developed that allows for a multi-criteria evaluation on different types of User Profiling and possible segmentation bases.

Practical Implications - The framework and model can support researchers and practitioners to determine which UML algorithms are appropriate for developing robust User Profiles and data-driven segments. The discovered profiles provide valuable insights for the UT M&C department to tailor marketing campaigns and improve strategic decision making.

Theoretical Implications - The framework and model contribute to literature regarding approaches and methodologies for UML and User Profiling in a marketing context. A two-stage clustering or hybrid user profiling approach can alleviate the drawbacks of one-stage clustering or solely using implicit and explicit user profiling. Discovery of micro-behaviours demonstrated that the proposed methods can generate profound insights and are indicative of a good performance by complete linkage and k-modes on a moderate sized and low dimensional symmetric binary dataset.

Value/Originality – Originality lies in the combination of complete linkage with the hamming distance, followed by the k-modes algorithm. To the best of the authors knowledge, this combination has not been used in academic literature, especially in education recruiting. Moreover, originality lies in including two-stage approaches for different types of data and data properties in the framework. The value of the model lies in including criteria for effective segmentation and different types of user profiling.

Keywords: Behavioural Targeting · Unsupervised Machine Learning · User Profiling · Categorical Data · Digital Marketing · Education Recruiting

Table of Contents

List of Abbreviations

AM	Applied Mathematics	HS	Health Sciences
AP	Applied Physics	IDE	Industrial Design Engineering
BE	Biomedical Engineering	IEM	Industrial Engineering and Management
BA	Business Administration	IT	Interaction Technology
BIT	Business Information Technology	IST	Internet Science and Technology
CE	Chemical Engineering	ME	Mechanical Engineering
CEM	Civil Engineering and Management	N	Nanotechnology
CS	Communication Studies	PSTS	Philosophy of Science Technology Society
CPS	Computer Science	P	Psychology
CME	Construction Management and Engineering	PA	Public Administration
EST	Educational Science and Technology	SE	Spatial Engineering
EE	Electrical Engineering	SET	Sustainable Energy Technology
ES	Embedded Systems	SC	Systems and Control
EEM	Environmental and Energy Management	TM	Technical Medicine
ES	European Studies	WT	Water Technology
GISEO	Geo Information Science and Earth Observation		

1. INTROCUTION

1.1 Background Research

Marketers are continuously challenged to understand consumer behaviour in order to improve an organization's market position. A key competitive advantage for today's organizations is the availability of large amounts of data for the purpose of segmenting a customer base, offering tailored services, and extracting meaningful information provided by various data sources. Customer (i.e., user) segmentation is one of the most central strategic issues in marketing. A fundamental task of segmentation is to group customers or users on the basis of similarities and develop specific marketing mixes or approaches per segment (Kotler, 2000). Tailoring an organisations offerings with the needs of a particular customer group enables the organization to gain a competitive advantage in the marketplace (Dolnicar 2008; Hiziroglu 2013). However, the success of targeted marketing efforts depend on the quality of the data-driven segments constructed. Today, organizations are confronted with rapid environmental changes such as technological developments and an increased audience fragmentation. The internet empowered consumers to gather quality information when planning to purchase new products and services. Therefore, organizations search for the most effective and efficient way to get their message in front of the right audience (Srimani & Srinivas, 2011). Moreover, organizations have been shifting their attention to generating online leads which refers to "an online visitor who registers, fills out a form, signs-up for, or downloads something on a website" (Mota et al., 2016, p. 134). Due to widespread internet use and advancements in consumer tracking technology, digital marketing can now be enhanced by Behavioural Targeting (Summers, Smith, & Reczek, 2016).

According to Srimani & Srinivas (2011) Behavioural Targeting (BT) is the ability to target users based on their behaviour on internet. Others define BT as an internet-based targeting strategy that uses several elements of a consumer's online behaviour to create a *user profile* which determines the content displayed to the specific individual (Lu, Zhao, & Xue, 2016; Summers et al., 2016). According to Summers et al. (2016) organizations can collect information of consumers by placing tracking technology (i.e., cookies) on their hard drive. This technology enables to collect browsing data, search history, media consumption, data from apps, purchases, click-through responses, e-mails, or social media (Boerman et al., 2017). A User Profile can be created from the data so that software is able to predict what could be appealing to a certain individual (Summers et al., 2016). According to the Internet Advertising Bureau (IAB), the economic value of BT in digital marketing include the following trends: (1) Digital Marketing in the EU generated €41.9 billion, with a growth rate of 12.3% in 2016. (2) BT is used in 66% of all digital advertising and contributes to 90% of digital advertising growth. (3) Data-driven marketing is over 500% more effective than marketing advertising that is not data-driven (IAB, 2017). These figures demonstrate the importance of leveraging customer data to gain a competitive advantage. Traditionally, segmentation was based on explicit information whereas BT utilizes implicit information or a combination of both types. BT techniques enable to distinguish individual differences in behaviour between two apparently similar customers. Traditional techniques often ignore such differences resulting in more heterogeneity within segments. Machine Learning (ML) can play a key role to gain insights from unstructured data. According to Bose and Mahapatra (2001) ML is "the study of computational methods to automate the process of knowledge acquisition from examples" (p. 212). ML can be divided into *unsupervised* and *supervised learning*. In *Unsupervised Machine Learning* (UML), no target variable is specified and only input data are provided (Larose, 2014). In contrast, *Supervised Machine Learning* (SML) algorithms are given a specific goal (e.g., target variable) for grouping data (Prasad, 2016). This paper focuses on UML which is commonly used for clustering and gaining insights from unstructured data.

1.2 Research Problem and Research Questions

Advancements in the Internet of Things, Neuroscience, Artificial Intelligence, and Data Mining have propelled the desire and collection of personal data for strategic decision making and personalisation (Chester, 2012). However, online customer data is considered to be one of the most underutilized sources of information. According to Subramaniam, Woo Tan, and Welge (2001) insights into behavioural characteristics and conversion patterns of users are often hidden or untapped by organizations. Similarly, according to Diapouli et al. (2017) organizations are often unable to gain meaningful insights out of data whereby a considerable amount of opportunities, resources, and marketing efforts are wasted. Moreover, the interpretability of data-driven segments continues to be an important research gap due to increasingly complex segmentation bases and a lack of guidance by literature (Dolnicar, 2009; Boratto et al., 2016). Additionally, there is a lack of understanding about the basics of data-driven segmentation methodologies among marketing departments (Dolnicar, 2009; Boratto et al., 2016). Key issues in methodological decisions for data-driven segmentation are determining the number of clusters and which algorithm should be chosen (Dolnicar, 2009). The majority of prior research focused on the accuracy, effectiveness, and efficiency of various Behavioural Targeting and Machine Learning techniques to improve online advertising. Additionally, the majority of research is limited to using one type of user data which is often explicit and metric. For instance, Yan et al. (2009) segmented users based on their responses to advertisements. Their experiment showed that click-through rates improved by 670 percent when using BT. Bhatnagar and Papatla (2001) segmented customers by using their search behaviour to present personalised ads. Targeting was based on the keywords a consumer entered in a search engine. Another technique used was monitoring the clickstream on advertisements to measure an ad's effectiveness in terms of click through ratios (Chen & Stallaert, 2014). Rindfleish (2003) focused on segment profiling based on geo-demographic data of students and how to use it to measure the potential of market segments in higher education. Yao et al. (2010) used Machine Learning to identify purchasing and spending amounts to generate customer profiles. Hence, numerous approaches and cases are available for numerical data but approaches for categorical or mixed types of data do not enjoy the same popularity. Moreover, none provided an outline of various Unsupervised Machine Learning approaches for User Profiling on categorical, numerical, and mixed types of data with respect to the characteristics of the dataset. Furthermore, prior research did not consider the different types of user profiling and the criteria that are essential for effective segmentation. Therefore, it is important to outline approaches in order to support researchers and practitioners to select appropriate methods and gain valuable insights out of data.

The *first objective* is to develop a methodology and a framework of Unsupervised Machine Learning algorithms for User Profiling with respect to important data properties. The *second objective* is to conduct a case study by utilizing the framework on data of University of Twente (UT) website visitors. The Marketing and Communications department (M&C) of the UT is among others responsible for monitoring the Higher Education market and developing student recruitment campaigns. A lot of data is available from the UT *CRM-system* and *Google Analytics*. Until now the M&C department was not able to find the right structure in their data and develop behavioural profiles. Moreover, the higher education market (HE) has to cope with increasing competition to recruit students. Marketing concepts which have been effective in business, are now needed by many universities looking to gain a competitive edge and gaining market share (Hemsley-Brown, & Oplatka, 2006). Changes in the HE market are, among others, caused by the increasing cost of education, globalization, or numerus fixus (Barber et al., 2013). Furthermore, Barber et al. (2013) argues that it is of increasing importance that "each university needs to be clear which market segments it wants to serve and how" (p. 5). Additionally,

potential applicants face complex challenges of narrowing down personal interests and motives into a single HE programme.

Hence, the M&C department can benefit from BT and ML techniques to develop more efficient and effective marketing campaigns. 10.435 students enrolled in 2017, including 79 different nationalities (Facts & Figures, 2018). However, Dutch students are the largest group of applicants and belong to the majority of website visitors (Facts & Figures, 2018). In order to develop the most accurate behavioural profiles the researcher specifies the target data to only include visitors interested in *Master* studies. To discover differences among behavioural profiles the selected data will consist out of two groups: (1) all website visitors interested in UT Master studies and (2) all Dutch website visitors interested in UT Master studies.

In brief, the *first objective* is to develop a framework of UML algorithms for User Profiling with respect to their requirements regarding data properties. The *second objective* is to discover high converting online behavioural profiles among Dutch website visitors interested in Master studies at the University of Twente (UT). The framework is aimed at supporting researchers and practitioners to determine which UML approach is most appropriate for developing robust user profiles. The discovered profiles can enable the UT M&C department to develop more effective and efficient marketing campaigns. The research questions are as follows:

1. *What is an appropriate framework for outlining UML Algorithms for User Profiling?*

The following sub-questions are addressed:
- What UML Algorithms are appropriate for categorical, numerical, or mixed data?
- What are their requirements regarding important data properties?

2. *What online behavioural profiles of Dutch website visitors interested in UT Master studies are most significant in terms of conversions?*

The following sub-questions are addressed:
- What UML Algorithm/similarity measure is appropriate for a symmetric binary dataset?
- What are the characteristics of different types of user data and customer attributes for Profiling?
- What type of User Profiling is appropriate for developing robust User Profiles for marketing purposes?
- What data mining/knowledge discovery process is appropriate?
- What segmentation criteria are essential for User Profiling?
- Are the discovered behavioural profiles among Dutch website visitors interested in Master studies consistent with the behavioural profiles of all website visitors interested in Master studies at the University of Twente?

1.4 Thesis Outline

The paper is organised in 5 chapters and structured as follows. The next chapter covers the theoretical framework whereby literature is reviewed on behavioural targeting, knowledge discovery, machine learning, customer segmentation, and user profiling. Chapter 3 outlines the research methodology based on the process of Knowledge Discovery in Databases (KDD). In chapter 4 the results are analysed and presented. Moreover, a comparison analysis is conducted to identify differences between All Visitors and Dutch Visitors interested in UT Master studies. Chapter 5 includes the discussion and conclusion, theoretical and practical implications, directions for future research, and research limitations.

2. THEORETICAL FRAMEWORK

A literature review is conducted on the core topics of this research which includes a definition of behaviour, Behavioural Targeting, Knowledge Discovery, Machine Learning, Segmentation Approaches, and User Profiling. Reviewing the core topics enables the researcher to develop a methodology and a framework including Unsupervised Machine Learning strategies for User Profiling with respect to the data properties. Additionally, a model can be developed for a multi-criteria evaluation on different types of User Profiling and customer attributes for effective segmentation. Relevant Literature of each subject is summarized and discussed briefly. An overview of the literature search strategy is given in appendix A.

2.1 Definition of Online Behaviour

Understanding the meaning of behaviour is essential in order to discover behavioural profiles. Behaviour can be explained as the manner of behaving or acting, and the action or reaction of systems and organisms under various circumstances. According to Cao (2014) behaviours can be recognized by the actions and mannerisms made by such organisms or systems in conjunction with their environment. Examples of more common terms are human behaviour, customer behaviour, or organizational behaviour. However, behaviour in the non-digital world is explicit and therefore it has been vastly studied from various aspects (Cao, 2014). Developments in computing technologies enabled a more social and digitalized life wherein behaviour becomes increasingly more complex as it includes the implicit form of digital information (Cao, 2014). For example, this may include the way an individual search for information or reacts to the digital environment. Behaviours documented in a digital format are often referred to as Behaviour Informatics or Behaviour Computing (Cao, 2014). Among others, behaviour informatics and behaviour computing consist of methodologies and techniques to represent, model, analyse, discover, and utilize human and organizational behaviours, virtual behaviours, behavioural relationships, and behavioural patterns (Cao & Yu, 2012; Cao, 2014). Furthermore, Cao (2010) refers to behaviour as “activities that present as actions, operations, events or sequences conducted by humans in a specific context and environment in either a virtual or physical organization” (p. 3069). Hence, behavioural patterns are an increasingly important asset to analyse and understand in order to disclose the implicit and explicit business value (Cao, 2014). A pattern can be described as “an expression in some language describing a subset of the data or a model applicable to the subset” (Fayyad et al., 1996). However, such patterns need to exceed a particular threshold in order to provide useful knowledge. Therefore, Fayyad et al. (1996) argues that the discovered patterns must be understandable and valid to some degree of certainty on new data that could provide meaningful information that benefits its users.

In brief, according to Cao (2010) behaviour can be defined as “activities that present as actions, operations, events or sequences conducted by humans in a specific context and environment in an either virtual or physical organization” (p. 3069). An example of behaviour in a digital form includes the actions manifested by visitors whilst surfing the UT website in order to acquire information. A combination of behaviours represent a behavioural pattern by which behavioural profiles can be described in this study. Techniques for data analysis are carefully selected and applied to the behavioural data of website visitors in order to extract meaningful information and develop behavioural profiles. Therefore, the following sections describe the fundamentals of such techniques that allows information to be extracted from raw data sources. In this study the raw data sources consist of behavioural data of UT website visitors.

2.2 Knowledge Discovery and Data Mining

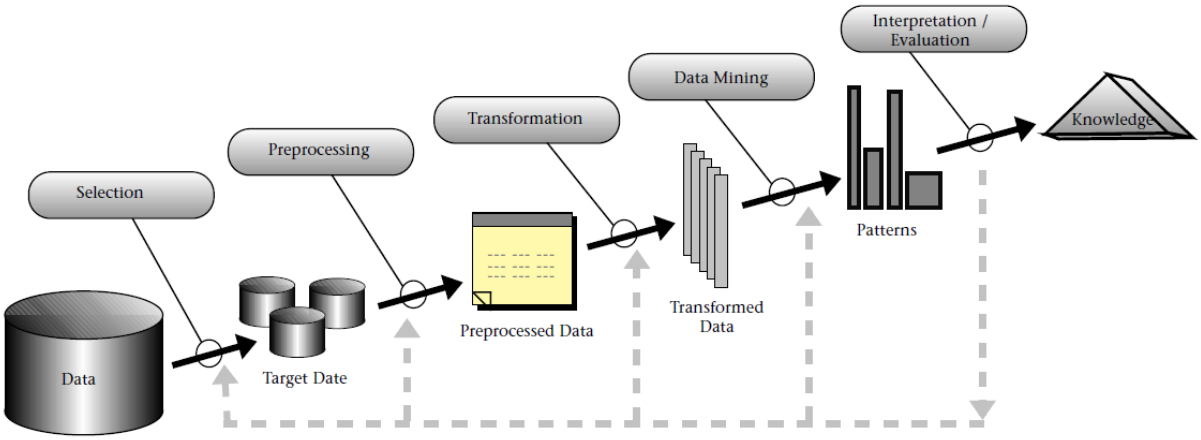
A key competitive advantage for today's organizations is to be able to explore data in order to understand customer behaviour, segmenting a customer base, offering tailored services, and gaining meaningful insights from data provided by various sources. Traditionally, researchers and practitioners used (statistical) surveys to study customer behaviour or relied on manual analysis and interpretation to discover knowledge (Romdhane, Fadhel, & Ayeb, 2010; Fayyad et al., 1996). However, advancements in information technology enabled organizations to generate large volumes of data as a result of monitoring business processes, user activity, website tracking, sensors, finance, human resources, and accounting (Assunção et al, 2015). Therefore, various data mining techniques have been developed in order to extract knowledge from data (Romdhane et al., 2010).

First, it is important to describe the relationship between the concepts of *Knowledge Discovery in Databases* (KDD) and *Data Mining*. Fayyad et al. (1996) defines KDD as "a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (p. 4). As mentioned previously, data refers to a set of facts and a pattern to "an expression in some language describing a subset of the data or a model applicable to the subset" (Fayyad et al., 1996, p. 5). *Data Mining* is referred to as the application of specific algorithms for extracting patterns from data (Fayyad et al., 1996) . A similar but more recent paper of Assunção et al. (2015) states that KDD is a process to extract non-obvious information and data mining refers to unveiling previously unknown patterns or interrelations among apparently unrelated attributes and datasets by utilizing methods from different areas, such as statistics and machine learning. These analytics consist of techniques including KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced visualization to support decision making (Assunção et al., 2015). It can be concluded that KDD can be regarded as the overall process of discovering useful knowledge from databases whereas data mining can be considered as a particular step within this process which is concerned with the application of algorithms.

2.2.1 Fundamentals of Knowledge Discovery in Databases

In order to extract knowledge from data the concept of *Knowledge Discovery in Databases* (KDD) was introduced by Fayyad et al. (1996). They distinguished two main categories of knowledge discovery goals, including *Verification* and *Discovery*. Verification refers to verifying a user's hypothesis whereas Discovery refers to the autonomous identification of patterns within data (Fayyad et al., 1996). Additionally, the Discovery category is divided into two sub-categories of *Prediction* and *Descriptive*. Prediction attempts to predict a future event or behaviour by using historical data. The descriptive sub-category aims to identify naturally occurring patterns in the dataset, creating management reports, and is concerned with modelling past behaviour (Fayyad et al., 1996; Assunção et al., 2015). Recently, *prescriptive* analytics emerged which assist users in decision making by determining actions and assessing their impact regarding business objectives, resources, and constraints (Assunção et al, 2015). Hence, this research is primarily concerned with discovery-oriented data mining in the descriptive sub-category. The KDD techniques have been widely applied in marketing, fraud detection, telecommunication, and manufacturing (Fayyad et al., 1996; Preeti et al., 2016). Hence, Knowledge Discovery is a research field concerned with the development of methods and techniques for making sense of data (Fayyad et al., 1996; Preeti et al., 2016). For example, a marketing application of KDD is to analyse business data to identify customer needs, distinct customer groups, or predict customer behaviour. The KDD process involves multiple iterative steps as depicted in Figure 1.

Figure 1. KDD-process (Fayyad, Shapiro, & Smyth, 1996; Preeti, Kalia, & Rani, 2016)



According to Fayyad et al. (1996) the first step is to determine the knowledge discovery goal and understanding the application domain. The second step is selecting the dataset or a subset of data on which discovery is conducted. Thirdly, the data is pre-processed if necessary, including data cleaning, removing noise, handling missing data fields, or accounting for unknown changes. Data transformation is the Fourth step and includes discovering useful attributes to represent the selected data depending on the research goal. One example is dimensionality reduction (e.g., Factor Analysis) to reduce the number of variables under consideration. The Fifth step is to match the Knowledge Discovery goals to specific data mining methods such as clustering, classification, and regression. Step six includes selecting an appropriate algorithm (e.g., for categorical data) and selecting methods to identify data patterns (e.g., distance measures). Step 7 is data mining which includes searching for patterns of interest in a representational and understandable form. Step eight is interpreting the discovered patterns, visualization of patterns, and possibly reconsidering step 1-7. Finally, step nine includes using the discovered knowledge directly or simply reporting it to interested parties (Fayyad et al., 1996). A comprehensive overview of various other data mining and knowledge discovery process models and their application areas are depicted in appendix B.

In summary, this research considers using the KDD-process because (1) the processes are similar, (2) KDD is widely applied in academic research and marketing, and (3) KDD is comprised out of more complete stages. Furthermore, the goal of this research is primarily concerned with discovery-oriented data mining and the descriptive sub-category which is used to identify naturally occurring patterns in data. In contrast, the prediction sub-category can be used in future studies for predicting customer behaviour with labelled data. Techniques for the descriptive and prediction sub-categories are outlined in the following sections.

2.2.2 Fundamentals of Data Mining

The data mining component of the KDD process involves iterative application of specific data mining methods. Generally, data mining methods consist of three primary algorithmic components including (1) model representation, (2) model evaluation, (3) and search (Fayyad et al., 1996). Respective to each of the three aspects: (1) refers to the language used to describe the discoverable patterns. It is important to understand the representational assumptions which might be inherent to the data mining method. (2) refers to statements of how well a model or pattern meets the knowledge discovery goals and (3) refers to parameter search and model search to fully optimize the data mining model. As mentioned, the primary goals of data mining are prediction and description. Related data mining methods can perform one or more of the following types of data modelling: Classification, Clustering, Regression, Association, Sequence Discovery, Summarization, Dependency Modelling, Deviation Detection, and Data

Visualization (Fayyad et al., 1996; Shaw et al., 2001; Ngai, Xio, & Chau, 2009). Furthermore, Data Mining involves selecting, exploring, and modelling large data sets to reveal unknown patterns and comprehensible information from large databases (Shaw et al., 2001). Big Data and Data Mining are therefore closely associated. Big Data is characterized by the three V's including Volume, Variety, and Velocity (McAfee & Brynjolfsson, 2012). However, Demchenko et al. (2013) has extended the three V's by including Veracity and Value. Volume refers to the data size, velocity to the data production and processing speed, variety to the distinct data types, veracity to the data validity in relation to its intended use, and value represents the worth derived from exploiting Big Data (Assunção et al, 2015). However, utilizing such analytics is still a labour intensive task because contemporary solutions for analytics are often based on appliances or software built for general purposes (Assunção et al, 2015). Hence, substantial effort is needed to tailor such solutions to the specific needs of the organisation or knowledge discovery goal.

In the field of information technology, data mining methods can be divided into two main categories of Machine Learning including *unsupervised* and *supervised* learning (Larose, 2014; Prasad, 2016; Walter & Bekker, 2017). The unsupervised method is associated with the descriptive sub-category of knowledge discovery as described by Fayyad et al. (1996). This type of Machine Learning aims to unveil naturally occurring patterns within the data without a target variable (Larose, 2014). The Supervised Machine Learning method relates to the prediction sub-category of knowledge discovery goals. The latter is given a specific target variable to classify certain events, objects, or attributes within the database to predict a future event based on historical data (Larose, 2014). Hence, the goal of this research is primarily concerned with Unsupervised Machine Learning. The following section provides a description of *Machine Learning* and its *Unsupervised* and *Supervised* Learning methods.

2.3 Machine Learning

The beginning of artificial intelligence (AI) in academic literature can be found around 1950 wherein Turing (1950) wrote the paper Computing Machinery and Intelligence. The topic received a lot of attention and particularly in recent years. Within AI, Machine Learning (ML) has become the technology of choice in achieving practical solutions (Jordan & Mitchell, 2015). They argue that the fast decrease in the cost of computational power and the availability of accumulating amounts of data are the two factors that drive the developments in ML. ML is currently on the top of the hype cycle which is characterized by extremely high expectations (Gartner, 2016). Hence, the expectations will drop significantly when a technology passes the top of the cycle. ML will become a mainstream application within the next three years if it proceeds through the hype cycle as expected. ML can play a key role in data mining applications to gain insights from unstructured data. According to Bose and Mahapatra (2001) ML is "the study of computational methods to automate the process of knowledge acquisition from examples" (p. 212). The goal of ML is to create algorithms which can learn or make predictions based on data and feedback. An important feature is that ML is not programmed to follow particular decision rules to create results, but rather, it has the capability of creating those rules by data and feedback (Jordan & Mitchell, 2015). ML techniques can be divided into two main categories of *unsupervised* and *supervised* learning (Larose, 2014; Prasad, 2016). The techniques and requirements related to both categories are described in the following subsections.

2.3.1 Unsupervised Machine Learning

In *unsupervised Machine Learning*, no target variable is specified and only input data are provided (Larose, 2014). *Clustering* and its variations are often referred to as Unsupervised Machine Learning (Larose, 2014; Prasad, 2016; Walter & Bekker, 2017). Clustering is used to discover the natural or arbitrary structural patterns in data determined by calculating the

distances between data entries. *Clustering* is a multivariate technique whose primary purpose is to group objects so that each object is similar to the other objects in the cluster and different from objects in all the other clusters (Larose, 2014; Prasad, 2016). Examples of applications of clustering analysis are understanding consumer behaviour by identifying homogeneous groups of customers, identifying new product opportunities by clustering products or brands, relationship identification, or for data reduction purposes. Clustering can be regarded as market segmentation which is one of the most central strategic issues in marketing (Dolnicar, 2002). The success of targeted marketing activities depend on the quality of the (data-driven) market segments constructed. Hence, a benefit of clustering lies in being able to tailor an organisations offerings with the needs of a particular customer group, in doing so, the organization gains a competitive advantage in the marketplace (Dolnicar 2008; Hiziroglu 2013). Important issues and requirements for clustering analysis are the research question being addressed, variables used to characterize objects, data type, data size, data dimensionality, distance measures, outlier detection, and the interpretability (Han, Kamber, & Pei, 2012; Larose, 2014).

The major fundamental clustering algorithms can be classified as: (1) *Hierarchical-based,* (2) *Partitioning-based,* (3) *Density-based,* (4) *Grid-based,* and (5) *Model-based* (Han et al., 2012; Fahad et al., 2014). In Density-based methods objects are separated based on their density, connectivity, and boundary (Fahad et al., 2014). Here, the density of objects is analysed to determine the functions of datasets that influence a particular object. In Grid-based methods the space of the data objects are separated into grids. In Model-based methods the fit between the data and a predefined mathematical model is optimized based on the assumption that the data includes a mixture of underling probability distributions (Fahad et al., 2014; Han et al., 2012). Model-based methods are able to automatically determine the number of clusters and taking outliers into account. Examples are Neural Networks such as Self-Organising Maps developed by Kohonen (1982).

For the sake of brevity, this study is limited to *Hierarchical-based* and *Partitioning-based* methods. These are discussed in more detail in the following sections to demonstrate their suitability in relationship to the goal of this research. Moreover, Dolnicar (2002) studied the standards of various clustering methods used in academic literature and found that the majority of segmentation applications (73%) either used hierarchical or partitioning methods.

2.3.2 Hierarchical Clustering

Hierarchical clustering methods are aimed at finding a structure in the data (i.e., a hierarchy) depending on the medium of proximity and are represented in a tree-like structure known as a dendrogram. Hierarchical clustering can be either *agglomerative* (i.e., bottom-up) or *divisive* (i.e., top-down). Agglomerative clustering initiates with one object for each cluster and reclusively merges it with two or more similar clusters (Fahad, 2014). A divisive variant operates in the opposite direction, wherein it initiates with the dataset as one cluster and reclusively separates objects to the most appropriate clusters (Fahad, 2014). However, drawbacks of hierarchical methods are that they cannot handle large datasets or high dimensionality well (Fahad, 2014; Pandove, Goel, & Rani, 2018). An advantage of hierarchical methods is that it is not required to specify the number of clusters a-priori. Furthermore, five agglomerative approaches exist including *Single Linkage, Complete Linkage, Average Linkage, Centroid's method, and Ward's method* (Fahad et al., 2014; Tamasauskas et al., 2012). According to Malhotra (2004) *Single linkage* combines two clusters with the smallest distance between objects and can be helpful to identify outliers but may depict snakelike "chains" clusters. *Complete Linkage* combines clusters with the smallest largest distance between objects and eliminates the chaining problem but is affected by outliers. *Average Linkage* combines two clusters with the smallest average distance between objects and is less affected by outliers. The *Centroid's Method* measures the smallest distance between cluster centroids and is less affected

by outliers. *Ward's Method* combines clusters so that the within cluster variance of the new cluster is as small as possible. It leads to equilibrated clusters, but it is easily distorted by outliers (Malhotra, 2004). More than half of the studies considered by Dolnicar (2002) used Ward's method for data-driven market segmentation.

2.3.3 Non-Hierarchical Clustering

Non-hierarchical clustering algorithms divide data objects into several partitions where each partition represents a cluster. Non-hierarchical methods are commonly used for handling large datasets because they are computationally less expensive (Fahad et al., 2014; Pandove, 2018). Non-hierarchical clustering can be *Hard* or *Soft* (Prasad, 2016). The basic methods typically adopt hard clustering known as *exclusive cluster separation* (Han et al., 2012). Here, each object must belong to exactly one group. In soft methods this requirement is relaxed by techniques such as fuzzy clustering. A requirement (or drawback) of non-hierarchical methods is that the number of clusters need to be specified beforehand so that initial seed points can be provided according to some practical, objective, or theoretical basis. However, non-hierarchical clustering methods are generally more robust against outliers. The k-means algorithm is one of the most prevalent in research (76%) (Dolnicar, 2002). In k-means, the centre is the average of all points representing the arithmetic mean (Fahad et al., 2014). K-modes replaces the means with modes (Huang, 1998). Other examples are k-medoids where objects near the centre represent the cluster (Fahad et al., 2014). However, most methods are *distance-based*. Distance measures are often used as a measure of similarity where higher values indicate greater dissimilarity between cases. These measures are calculated across the entire set of clustering variables which allow for the grouping of observations and their comparison to each other. However, distance measures should be chosen in accordance with the data format. Various distance measures available, with Euclidian distance being the most popular similarity measure in academic literature (Dolnicar, 2002).

In brief, non-hierarchical methods are preferred for large datasets and are more robust against outliers. Hierarchical clustering is preferred when more than one clustering solution is of interest or the sample size is moderate. A key step in applying such methods is to select an appropriate similarity measure based on the data type to calculate the distance between objects. For non-hierarchical methods it is required to specify the number of clusters. The data in this study is categorical. Specifically, it concerns a symmetric binary dataset. Hence, it is important to address the issues stated above in the following sections to select the most appropriate *algorithms, distance measures,* and approaches to *determine the number of clusters* for a symmetric binary dataset for unsupervised machine learning.

2.3.4 Binary Data: Algorithms, Similarity Measures, and Data Properties

Determining the similarity measure to calculate the distance between objects is a key step for clustering analysis. Similarity measures for continuous data are relativity well-understood and widely available but for categorical data it is not as straight forward (Boriah, 2008). In contrast to continuous data, categorical data is deficient of default ordering relationships on the attribute values which make the task of developing distance measures and clustering algorithms for categorical data more challenging (Alamuri, Surampudi, & Negi, 2014). A distinctive characteristic of data mining applications is that it deals with large, complex, or high dimensional datasets. Datasets can include millions of objects and hundreds of attributes. Attributes can be divided into metric (i.e., quantitative) or nonmetric (i.e., qualitative). Nonmetric measurement scales are *nominal* (e.g., gender), *ordinal* (e.g., education level), and *binary* (e.g., yes/no) whereas metric measurement scales are *interval* (e.g., temperature) and *ratio* (e.g., weight) (Huang 1998; Prasad, 2016). Hence, ML algorithms are therefore required to be scalable and capable of handling different types of attributes. Interesting clustering

algorithms are those who can handle large datasets of numeric or categorical variables because these types of data are most frequently present in real world data (Dolnicar, 2002). However, most clustering algorithms can either handle large data sets but are limited to numeric attributes or they are able to handle both types of data but are inefficient at handling large datasets (Fahad et al., 2014).

For *non-hierarchical clustering*, MacQueen (1967) introduced the *k-means* algorithm which can efficiently handle large datasets and is therefore well suited for data mining tasks. In the k-means algorithm the centre is the average of all points representing the arithmetic mean (Fahad et al., 2014). It iteratively searches the cluster centres and updates the memberships of objects to minimise the within cluster sum of squares (WCSS) using the *(squared) Euclidean distance* measure. A drawback is that k-means only works efficiently on numerical data (MacQueen, 1967; Fahad et al., 2014). Huang (1998) introduced the *k-modes* non-hierarchical algorithm which is suitable for clustering large categorical datasets. The key differences are that k-modes uses a *simple matching dissimilarity measure* (i.e., hamming distance) instead of Euclidean distance, replaces the means of clusters with modes, and uses a frequency-based method to update cluster modes (Huang, 1998). Using the modes of clusters makes more sense for categorical data than using means or averages. The k-modes dissimilarity measure is defined by the total mismatches of corresponding attribute categories of the two objects (Huang, 1998). Hence, the smaller the amount of mismatches the higher the similarity between objects. Furthermore, k-modes is faster compared to k-means because it converges in less iterations (Huang, 1998). A similar algorithm is *k-medoids* introduced by Park and Jun (2009) wherein medoids are considered instead of centroids or modes. It is based on the most centrally located object within a cluster and therefore less sensitive to outliers (Park & Jun, 2009). Hence, k-medoids is suitable for categorical data and handling outliers (i.e., noise) but it does not handle large datasets efficiently (Fahad et al., 2014).

The non-hierarchical methods mentioned above are most suitable to either handle numerical or categorical attributes. However, objects encountered in real world databases are often *mixed-types of data*. Huang (1998) integrated the k-means and k-modes algorithms and introduced the *k-prototypes* algorithm that can be used to cluster mixed-type objects and is capable to handle large datasets and high dimensionality. The algorithm includes the squared Euclidean distance measure for numeric attributes and the simple matching dissimilarity measure for categorical attributes (Huang, 1998). A certain weight is used to avoid favouring a type of attribute whereby the researcher's knowledge about the data is an important factor.

For *hierarchical clustering* various algorithms are available in literature. Guha, Rastogi, and Shim (1998) introduced and applied the hierarchical algorithm *CURE* for clustering large datasets. The algorithm considers the scattered points as representatives to capture the shape and extent of the cluster (Guha et al., 1998). The closest pair of representative points are merged at each step to generate the clusters. According to Guha (1998) and Fahad et al. (2014) it can not only handle large datasets but also high dimensionality and it is more robust against noise because shrinking the scattered points toward the mean reduces sensitivity to outliers. However, it is applicable on numerical data only (Fahad et al. 2014). Karypis, Han, and Kumar (1999) introduced and applied the hierarchical algorithm *Chameleon* which is based on dynamic modelling. A key feature is that it considers the interconnectivity and closeness in identifying the most similar pair of clusters (Karypis, 1999). Hence, two clusters are merged when the interconnectivity and proximity (closeness) between clusters is high compared to the within cluster interconnectivity and closeness of objects. Karypis et al. (1999) states that as long as a similarity matrix can be provided, the dynamic modelling of clusters in the Chameleon algorithm is applicable to all types of data. Guha et al. (2000) introduced the *ROCK* algorithm which is applicable to both numerical and categorical variables (Guha et al., 2000; Fahad et al., 2014). As argued in Guha et al. (2000) the ROCK algorithm uses a *links-based measure* and

not a distance-based measure as a basis to merge neighbouring data points to create clusters. While the ROCK algorithm is capable of handling large datasets, it is less efficient at handling high dimensionality or noise (Guha et al., 2000; Fahad et al. 2014).

A *binary* dataset is considered in this study whose values can indicate an attributes absence (0) or presence (1). Nominal scaled variables can only be allocated to different classes but cannot be ordered or measured like metric variables. Hence, the (dis)similarity or distance among two categorical attributes is proportional to the number of characteristics in which they match. Binary attributes can be *symmetric* or *asymmetric* (Ordonez, 2013). *Symmetric* binary data is when the outcomes are of equal importance and have assigned equal weight when calculating the similarity. A match of 0/0 or 1/1 are equally important. In contrast, matches of *asymmetric* binary attributes are not equally important (Ordonez, 2013). In this study, the matches are of equal importance and thus symmetric. For instance, it is of equal importance to consider visitors who manifested a particular behaviour (1/1) and visitors who did not (0/0) to discover accurate behavioural profiles. In contrast, a positive or negative result of a medical test might not be of equal importance. Hence, it is important to briefly discuss appropriate combinations of hierarchical clustering methods and distance measures for a symmetric binary dataset. Boriah et al. (2014) studied which similarity measures could be recommend and concluded there is no best performing similarity measure. However, for symmetric and asymmetric binary data Tamasauskas, Sakalauskas, & Kriksciuniene (2012) evaluated the performance of ten different hierarchical clustering methods by experimenting with ten different similarity measures in terms of accuracy. Similarity measures including the hamming distance, dmatch, dsqmatch, rogers and tanimoto, and sokal and sneath1 were tested on symmetric binary data. Djaccard, Dice, Russell and Rao, Bray and Curtis, and Kulcynski1 were tested on asymmetric binary data (Tamasauskas et al., 2012). Furthermore, the study included the hierarchical methods of single linkage, complete linkage, average linkage, centroid's method, density linkage, flexible-beta, McQuitty's, median, two-stage density linkage, and Ward's method. Performance evaluation revealed that the best methods are complete linkage, flexible-beta, and Ward's method (Tamasauskas et al., 2012). *Complete Linkage* performed best among all *symmetric* distance measures (Tamasauskas et al., 2012). An overview of the findings of Tamasauskas et al. (2012) is depicted in Appendix F.

In addition to hierarchical and non-hierarchical methods the *model-based* method is often used in academic literature for clustering. Dolnicar (2002) and Fahad et al. (2014) mentioned Neural Networks became a more prevalent application in literature for clustering solutions. According to Santana et al. (2017) the Self-Organising Maps (SOMs) algorithm introduced by Kohonen (1998) is the most used type of neural network. SOMs can provide models for clustering, classification, and forecasting (Sathya, & Abraham, 2013). The goal of SOMs is to convert an input signal (high dimensional) into a simpler discrete map (Larose, 2014). Additionally, it used for data visualization or dimensionality reduction purposes (Kohonen, 2013). SOMs structures output nodes into clusters of nodes where nodes in closer proximity are more similar than to other nodes that are further apart (Larose, 2014; Kohonen 2013). SOMs are less sensitive to initialization and it is not required to specify the number of clusters a priori (Murray, Agard, & Barajas, 2017). However, while SOMs is capable of handling high dimensionality, it is less robust against noise (Fahad et al., 2014). Another drawback of SOMs is that it is computationally expensive when handling large datasets (Murray et al., 2017). Moreover, SOMs was developed to cluster real-valued data whereby the range of variation allowed by Euclidean distance cannot be matched by binary measures (Lourenco et al., 2004). They concluded it is less appropriate to apply binary similarity measures when using SOMs and learning other data types remains a challenge. However, Santana et al. (2017) proposed an modified SOM for improved binary or categorical clustering. Results indicated that the modified SOM delivered more robust results compared to other SOM variants for binary data.

However, non-hierarchical clustering requires to specify the number of clusters a priori. When the number of clusters are not determined properly, it will significantly impact the results and mislead interpretations in data-driven market segmentation. The next section is aimed at proposing the solution as well as taking into account the sample size.

2.3.5 Two-Stage Clustering and Data Size

Determining the number of clusters a priori most strongly influences the clustering solutions. The problem of selecting the number of clusters is one of the oldest unsolved problems in clustering analysis (as cited in Dolnicar, 2002). One of the first approaches were suggested by Milligan (1981) and Milligan & Cooper (1985) which are based on an internal index comparison. However, a two-stage clustering methodology was proposed by Punj and Stewart (1983) wherein they recommended to identify clusters by first using Ward's method or average linkage (i.e., hierarchical clustering) followed by non-hierarchical clustering for cluster refinement. They concluded a two-stage approach yields better results than solely using a hierarchical or non-hierarchical approach. Mazanec and Strasser (2000) adopted a two-stage approach of hierarchical and non-hierarchical clustering and drew similar conclusions. Kuo, Ho, and Hu (2002) modified the two-stage approach and proposed to use self-organising maps to determine the number of clusters followed by the k-means algorithm. They concluded their modified two-stage method provided good solutions for determining the initial segments and observed a reduced number of misclassifications compared to conventional methods. Hence, determining the number of clusters by hierarchical clustering before applying a non-hierarchical procedure might be an advisable approach for this study.

Hierarchical clustering methods are computationally expensive and slow when handling large datasets or high dimensionality (Fahad et al., 2014). Therefore, literature is reviewed in order to provide some indications on what data size could be referred to as large or small. Generally, non-hierarchical methods have superior performance on large data sets whereas the performance of hierarchical methods decreased as the number of observations increased (Zhao & Karypis, 2002; Abbas, 2008). Dolnicar (2002) studied the standards of clustering analysis in academic literature for data-driven market segmentation and found that the smallest data size contained only 10 objects, the largest 20.000 objects, and the average size was 700. In case of hierarchical clustering methods the data sizes contained 530 observations on average and for partitioning methods 927. The number of variables in the datasets ranged between10 and 66 variables, with a mean number of 17 variables (Dolnicar, 2002; Dolnicar, 2003). Therefore, one could potentially regard 10 variables as low dimensionality and more than 10 variables as high dimensionality. Other studies have applied hierarchical clustering methods on varying data sizes. For instance, Abbas (2008) evaluated the performance of hierarchical and non-hierarchical clustering methods on data sizes of 4000 and 36000 with varying dimensionality and numbers of clusters. Results indicated that hierarchical clustering performed best on the smaller dataset with low dimensionality. Therefore, a data size of less than 4000 observations could potentially be considered as being small enough for hierarchical clustering and its computation time. Datasets with more than 4000 observations could be considered as large and potentially less suitable for hierarchical clustering methods except for the Chameleon, ROCK, and CURE algorithms. Due to a lack of rules regarding the data size, the only recommendation that could be given is to question if the dimensionality is not too high for the number of cases to be grouped (Dolnicar, 2002; Dolnicar, 2003). One approach to determine the minimum data size is to include no less than 2^k cases (k = number of variables), and preferably $5*2^k$ (Dolnicar, 2002). This study considers 10 behavioural attributes. Hence, the sample size should be at least between 1024 and 5120 observations according to the suggested recommendation.

In summary, for this study a combination approach is advisable using a hierarchical approach followed by a non-hierarchical approach. Hierarchical clustering is applicable when more than

one clustering solution is of interest or the sample size is moderate. The number of clusters is determined by hierarchical clustering and a non-hierarchical procedure then clusters all observations using the determined number of clusters or initial seed points to provide more accurate cluster memberships. The best performing hierarchical clustering method was *complete linkage* in combination with the *hamming distance* for moderate sized datasets, low dimensionality, and symmetric binary data. For non-hierarchical clustering the most appropriate algorithm is *k-modes* because it is specifically developed to handle categorical datasets and it is based on the simple matching dissimilarity measure (i.e., hamming distance). Lastly, no previous studies have been encountered in the field of higher education marketing and student recruitment that applied the combination of methods as proposed in this research.

2.3.6 Supervised Machine Learning

Supervised Machine Learning algorithms are given a specific goal (e.g., target variable) for grouping data (Larose, 2014; Prasad, 2016; Walter & Bekker, 2017). *Prediction* and *classification* are often regarded as Supervised Learning. In supervised learning the purpose is to learn from input variables whereby the correct values are provided by a supervisor (Walter & Bekker, 2017). Examples of classification techniques include: *Neural Network (SOMs), K-nearest neighbour, Decision Trees, Support Vector Machines, Bayesian, and naïve bayes* (Ngai, Xio, & Chau, 2009; Larose, 2014; Walter & Bekker, 2017). Classification is a type of prediction that partitions data into categorical variables. A well-known technique is the Decision Tree which makes use of recursive partitioning to divide the objects by a data-driven threshold for each variable in multiple levels (Chorianopoulos, 2016). Hence, a classification technique can be used to allocate observations to various pre-determined segments.

2.4 Behavioural Targeting

According to Srimani & Srinivas (2011) BT is the ability to target users based on their behaviour on internet. Moreover, BT can be defined as an internet-based targeting strategy that uses several elements of a consumer's online behaviour to create a *user profile* which determines the content displayed to the specific individual (Lu et al., 2016; Summers et al., 2016). In addition, BT techniques for online advertising is referred to as Online Behavioural Advertising (OBA). Boerman, Kruikemeier, and Borgesius (2017) define OBA as "the practice of monitoring peoples online behaviour and using the collected information to show people individually targeted advertisements" (p. 2). Hence, it can be concluded that BT is based on past individual-level (online) behaviour to determine a user's interest and accurately target potential consumers with tailored content. According to Summers et al. (2016) organizations are able to collect information of consumers by placing tracking technology (i.e., cookies) on their hard drive, enabling them to collect a visitor's viewing and clicking patterns, searches, conversions, or social media use. Data from online behaviour can consist of web browsing data, search history, media consumption (e.g., photos or videos), data from apps, purchases (i.e., conversions), click-through responses to ads, and communications such as e-mails or social media posts (Boerman et al., 2017). A user profile can be created from the data so that software is able to predict what could be appealing to a certain individual (Summers et al., 2016).

Different kinds of Behavioural Targeting (BT) techniques exist that serve different marketing purposes. Major categories are *Contextual BT, Onsite BT, Ad Networks BT*. Contextual Targeting (CT) aims to deliver online ads to a user based on the web content that is being viewed and aims to target consumers at the right time in a specific context (Lu et al., 2016). In contrast, BT aims to identify consumers who are more likely to be interested in the content presented, that is, the right audience (Lu et al., 2016). Furthermore, Lu et al. (2016) found that combining BT and CT has a positive interaction effect on a consumers conversion behaviour. Related types of targeting include retargeting, IP-based geo-targeting, explicit profile data targeting, and search targeting (Lambrecht, & Tucker 2013; Lu et al., 2016).

Additionally, two different types are distinguished namely Ad Networks BT and OnSite BT (Srimani & Srinivas, 2011). Ad networks refers to a company that serves advertisements on thousands of websites which enables them to collect data across various websites and ads (Boerman et al., 2017; Srimani & Srinivas, 2011). Onsite BT is aimed to improve a visitors experience on a single online property, such as a website (Srimani & Srinivas, 2011). An appropriate BT method can be selected, depending on the business goals, context, and information systems available. Traditional targeting techniques and BT are different in two ways. First, BT is the ability to target users based on data-driven segmentation of individual-level behaviour on internet whereas traditional targeting techniques are based on common sense segmentation of markets using explicit information related to geo-demographics, psychographics, or social identities under the assumption that these groups share certain characteristics (Summers et al., 2016). Secondly, content presented by BT techniques is more person specific whereas traditional targeting techniques present similar content or ads to all visitors (Summers et al., 2016). For example, segmentation done by country can result in more heterogeneity within segments whereas segmentation based on individual behaviour (e.g., interests and needs) can result in more homogeneity within segments.Various studies suggested that using BT generates more conversions and revenue compared to instances where BT was not used. Chen and Stallaert (2014) found that conversion rates on behaviourally targeted content was more than twice as high compared to traditional targeted content. Similar results of Goldfarb and Tucker (2011) indicated that users were less likely to convert after viewing content that was not behaviourally targeted. Yan et al. (2009) segmented users based on their browsing and search behaviour and compared advertisement responses across segments. Their experiment showed that click-through rates improved by 670 percent when using BT techniques. Similar results were presented by Bhatnagar and Papatla (2001) who segmented customers by using their search behaviour to present personalised ads. Targeting was based on the keywords a consumer entered in a search engine. Another technique used was monitoring the clickstream on advertisements to measure and ad's effectiveness in terms click through ratios (Chen & Stallaert, 2014).

The success of behavioural targeting activities depend on the quality of the (data-driven) segments identified. Additional information is required for a good understanding on how segmentation and user profiling are ought to be done. The following sections outline common approaches to segmentation and user profiling.

2.4 Segmentation Approaches

Segmentation is one of the most central strategic issues in marketing. A fundamental task of segmentation is to group customers on the basis of similarities and develop specific marketing mixes or approaches per segment (Kotler, 2000). Being able to tailor an organisations offerings with the needs of a particular customer group enables the organization to gain a competitive advantage in the marketplace (Dolnicar 2008; Hiziroglu 2013). Segmentation results should be simple to interpret while the accuracy of the segments is as high as possible. All segmentation approaches can be classified into two categories. On one hand there is the *common-sense* (Dolnicar, 2004) or *a priori* (Saia et al. 2016) approach and on the other hand the *post-hoc* approach, also known as *a posteriori* or *data-driven* (Dolnicar, 2004; Boratto et al., 2016). Common-sense is based on a simple property such as country which is used to segment users. This technique generates segments that are easy to understand and can be generated at a low cost (Boratto et al., 2016). However, this approach is trivial and runs the risk of superficial or generic segments. The post-hoc (i.e., data-driven/a posteriori) approach combines a set of attributes in order to create user segments (Boratto et al., 2016). Users are grouped based on data-driven similarities among multiple attributes. The post-hoc approaches provide more accurate segments (Dolnicar, 2004). However, due to a more complex segmentation base the

problem of properly understanding the results arises (Boratto et al., 2016). This is caused by a lack of guidance on how to understand the results of more complex segmentation approaches (Boratto et al., 2016). Easily understandable approaches generate ineffective segments while the complex approaches are accurate but not easy to use in practice. In order to address the shortcomings of common sense and data-driven approaches Dolnicar (2004) proposed a systematics resulting in a *hybrid* approach. The systematics leads to combining the aforementioned approaches as follows: Common-Sense/Common-Sense, Data-Driven/Common Sense, Common Sense/Data-Driven, and Data-Driven/Data-Driven segmentation (Dolnicar, 2004). However, the systematics do not include three-step approaches as well as simultaneous combinations of data-driven and common sense approaches (Dolnicar, 2004). In a study conducted a few years later, Dolnicar (2009) concluded that 65 per cent of the study subjects (Marketing Managers) have difficulties understanding a data-driven segmentation *solution.* Similarly, Boratto et al. (2016) argued that the understand-ability and interpretability of the segments continued to be an important research gap. The researcher could refer to this issue as the *managerial usefulness* of the results of a segmentation approach. For instance, the managerial usefulness of a user segment is higher when the results are easy to understand while maintaining a high match (segment quality) between the needs (i.e., segments) and offerings (i.e., organization). Furthermore, Dolnicar (2009) concluded that a large proportion of marketing managers lacked a fundamental understanding about data-driven market segmentation *methodologies*. Key issues in methodological decisions were determining the number of clusters, selecting the distance measure, and which algorithm should be chosen (Dolnicar, 2009).

In brief, there are three approaches to segmentation: Common Sense (a priori), Data-Driven (a posteriori), and hybrid. Common sense generates segments that are easy to understand but less accurate. Data-driven segmentation leads to segments that are more accurate but difficult to interpret. The hybrid approach includes combinations of segmentation approaches and alleviates the shortcomings of solely using either type. Furthermore, the approaches fail to acknowledge how different types of user data and evaluation criteria affect the managerial usefulness of the segmentation results. Hence, the following sections outline user profiling approaches based on different types of user data and criteria for effective user profiling.

2.5 Types of User Profiling

User profiling can be referred to as the process of gathering information specific to each visitor either explicitly or implicitly (Eirinaki &Vazirgiannis, 2003). A user profile generally includes a visitors demographic information, interests, or even their behaviour (Eirinaki &Vazirgiannis, 2003). The collected information can be used to personalize a website, ads, or various marketing efforts to a specific individual's interests. Poo, Chng, and Goh (2003) discussed various user profiling approaches and information filtering techniques. There are two types of user profiling namely, *static profiling* and *dynamic profiling,* and two kinds of information filtering namely, *Content-based filtering* and *Collaborative filtering* (Poo et al., 2003; Cufoglu, 2014).

Static profiling analyses a user's static and predictable attributes. Static information usually comes from the users themselves such as conducting online registrations or ratings (Poo et al., 2003). However, a static profile degrades in quality over time as the users interests changes (Poo et al., 2003). This may result in a more subjective view that not accurately reflects the interests of other users with similar interests. *Dynamic* profiling is the process of analysing a user's activities or actions to determine a user's interests (Poo et al., 2003). This can be referred to as behavioural profiling. This method provides a more objective and accurate representation of users interests.

Content based filtering compares the contents of items associated with a user profile and selects those documents whose contents best match the contents of another user profile (Poo et al., 2003 ; Cufoglu, 2014). This technique requires users to provide explicit feedback to the

system (e.g., ratings). This can be an issue as some users are reluctant to (voluntarily) provide such feedback. Hence, implicit user information is needed to address this problem. *Collaborative* filtering organizes users with similar interests into groups (Poo et al., 2003; Cufoglu, 2014). This is commonly done by clustering the users into different profiles. However, the effectiveness of this approach depends on how well the clustered profiles reflect the users interests (Poo et al., 2003). In order to alleviate these drawbacks Poo et al. (2003) proposed a *hybrid* user profiling system by combining the aforementioned concepts resulting in four user profiling approaches namely, static content profiling, static collaborative profiling, dynamic content profiling, and dynamic collaborative profiling. The model of Poo et al. (2003) is depicted in appendix D and user profiling methods of Cufoglu (2014) in appendix E.

Cogfoglu (2014), Kanoje,Girase, & Mukopadhyay (2014), and Khosrow-Pour (2009) referred to the static and dynamic profiling strategies based on the nature of information namely, *Explicit User Information* and *Implicit User Information*. *Explicit user profiling* refers to the static profiling paradigm whereby the interests of a user is known once a user provides the information. However, there are various issues to consider when using explicit user information. According to Schiaffino and Amandi (2009), users may be reluctant or unwilling to provide such information. Secondly, users may not always provide the truth when completing some kind of form about themselves. Thirdly, when users are willing to provide information, some might be less able to accurately express their interests and needs. Hence, a more accurate method is to obtain *implicit* user information by observing the users interactions with the underlying application, tracking these actions, and discovering patterns by some data mining technique (Schiaffino & Amandi, 2009). *Implicit user profiling* relates to the dynamic profiling paradigm. This type of data provides a more objective and accurate view on a user's interests. However, implicit user profiling often includes a more complex segmentation base which makes it less easy to understand. *Hybrid User Profiling* overcomes the drawbacks of explicit and implicit profiling by combining the two methods (Khosrow-pour, 2009, p. 2757). Hence, user profiles would reflect more accurate and realistic preferences and interests of users. It works by first considering explicit user data which is then updated and supported by the implicit user data or the method can be reversed. An overview of the User profiling Concepts discussed by Kosrow-Pour (2009) are presented in appendix C.

In brief, there are three types of user profiling which indicate that the nature of the collected information is important for obtaining meaningful and accurate user profiles. Static user profiling is based on obtaining *explicit* user information whereas dynamic user profiling is based on *implicit* user information. The hybrid approach combines both types of profiling to address their shortcomings in order to create superior user profiles. In order to make recommendations to groups, the individual profiles can be aggregated to obtain group profiles. This study initially obtains explicit data from the CRM database (e.g., Study Programme, Country, Study level). Secondly, implicit information is obtained from Google Analytics (e.g., PDF-downloads). However, user profiles can consist out of various types of customer attributes which are outlined in the following sections.

2.5.1 Segmentation Bases for User Profiling

There are various customer attributes that can be used for user profiling. However, it is important to recognize different categories to which these attributes are related in order to understand the information they yield. Utilizing various customer attributes may yield a different image of users and subsequently their user profiles. Profiling can be based on the following major characteristics, including: *Geographic, Demographic, psychographic, behavioural, Propensity-based, and Value-based* (Goyat, 2011; Chen & Stallaert, 2014; Tsiptsis & Chorianopoulos, 2011; Hiziroglu, 2013).

Geographic attributes allow to segment consumers based on location. Variables that are often used are region, population density, and climate (Goyat, 2011). *Demographic* attributes segments consumers according to their age, gender, education, family size, family life cycle, income, ethnicity, religion, occupation, or social class (Goyat et al., 2011). *Psychographic* attributes allows to segment consumers according to their interests, activities, opinions, values, or attitudes (Goyat, 2011; Tsiptsis & Chorianopoulos, 2011). Lastly, *behavioural* segmentation is based on the actual customer behaviour towards products and services including their needs and interests. Customers can be divided according to their identified behavioural and usage patterns. This type of segmentation is typically used to develop personalized offerings (Tsiptsis & Chorianopoulos, 2011). Examples are benefits sought, interests, preferences, intentions, brand loyalty, user status, or readiness to buy (Goyat, 2011; Jadczakova, 2013). *Propensity-based* attributes allow for the grouping of customer according to their propensity scores such as churn scores or cross-selling scores (Tsiptsis & Chorianopoulos, 2011). These kind of attributes can often be combined with other segmentation schemes for improved targeted marketing actions. Lastly, *value-based* segmentation groups customers according to their value. It can be used to identify the most valuable customers, to track value, and how value changes over time, for differentiation in service strategies, and optimization of resource allocation for marketing activities (Tsiptsis & Chorianopoulos, 2011). In addition, it is important to describe the main data sources that allow to extract such information for behavioural profiling.

2.5.2 Data Sources

Various data sources can be considered for extracting customer attributes and creating meaningful behavioural profiles. According to Araya, Silva, and Weber (2004) there are three major categories of data sources for mining web usage behaviour namely, web data, business data, and meta data. *Web data* is generated by a visitors actions and interactions on a website and stored in log files, cookies, and queries (Araya et al., 2004). *Business data* is data generated by the CRM systems of the respective business which often includes geo-demographics, product information, and other explicit data. According to Blackboard (2014) about 55% of higher education institutions do not use a CRM system for marketing or recruitment purposes. Moreover, *Meta data* describes the content and structure of the website. The structure is provided by the home page, links between pages, or navigational structure. The content is represented by a vector space model (Araya et al., 2004). In this study, web data is obtained by Google Analytics and Business Data is obtained by the UT's CRM-system. According to Singal, Kohli, and Sharma (2014) the two major approaches for gathering data for website analysis are *log files* and *page tagging*. Log files record the user interactions with the website such as page views and conversions. Page tagging are tags inserted in an existing HTML source code of a website. These page tags allow to track and analyse the behaviour of visitors whilst surfing the website. Examples of page tagging tools are Google Tag Manager and Google Analytics. However, this study requires behavioural attributes in order to discover behavioural profiles among website visitors of the University of Twente. Therefore, literature regarding behavioural attributes is reviewed in order to gain insights into what attributes might be appropriate to consider for this research.

2.5.3 Behavioural Attributes

Numerous academic studies are dedicated to behavioural attributes and user profiling in combination with various methods for data analysis. This study takes into account the assumption of the segmentation theory where groups of customers with similar behaviours and needs are likely to demonstrate a homogenous response to marketing activities (Tsai & Chiu, 2004). Examples of behavioural attributes encountered in literature are mentioned below (Goyat, 2011; Pandey et al., 2011 Tsiptsis & Chorianopoulos, 201; Gutwirth 2012; Baranowska, 2014).

Pages views	Number of visits	Date
Search Queries	Sequence of behaviour	Device
Ads clicked	Time spent on page	Benefit sought
Referring URL	Operating System	Product offers viewed
Location	In-text semantics	Social Media Channels
Purchasing activity	Campaigns	Conversion ratio
Clickstream	Navigational behaviour	Average time on page

2.6 Framework for User Profiling based on Unsupervised Machine Learning

A framework is proposed to visualize User Profiling strategies based on unsupervised machine learning and the requirements and characteristics of the dataset. The framework is based on literature discussed in section 2.3. Selecting a particular algorithm for unsupervised machine learning problems is highly dependent on the data type, data size, and data dimensionality. These data properties have a significant effect on the quality and efficiency of the clustering procedure and solution (Fahad et al., 2014, Pandove et al., 2018, Dolnicar., 2002). For instance, when analysing a large numerical dataset one might apply k-means and for categorical data k-modes. Additionally, a dataset containing numerous attributes is referred to as being high dimensional. However, only a limited number of algorithms is capable of handling high dimensionality. The main issue of high dimensionality is that objects appear to be alike due to a loss of meaningful differentiation between similar and dissimilar objects and the discriminative power of the similarity measure (Assent, 2012). There is a wide variety in the dimensionality and data size used in academic literature. Dimensionality can range from as little as ten attributes to thousands of attributes in domains such as molecular biology (e.g., Kailing et al., 2003). Similarly, a data size can range from a few objects to millions of objects. According Asset (2012) no standards or rules exist in literature which indicate what can be considered as a high dimensional dataset. Similarly, there is a lack of rules regarding the data size and what can be considered as small or large (Dolnicar, 2002).

However, Dolnicar (2002) studied the standards of clustering analysis in academic literature for data-driven market segmentation and found that the smallest data size contained only 10 objects, the largest 20.000 objects, and the average size was 700. The number of variables in the datasets ranged between 66 and 10 variables, with a mean number of 17 variables (Dolnicar, 2002; Dolnicar, 2003). Therefore, one could potentially regard 10 variables as low dimensionality and more than 10 variables as high dimensionality. Additionally, Abbas (2008) evaluated the performance of hierarchical and non-hierarchical clustering methods on data sizes of 4000 and 36000 with varying dimensionality and numbers of clusters. Results indicated that hierarchical clustering performed best on the smaller dataset with low dimensionality. Therefore, a data size of less than 4000 could potentially be considered small enough for hierarchical clustering and its computation time and interpretability. Datasets with more than 4000 observations can be considered as large and potentially less suitable for hierarchical clustering methods except for the Chameleon, ROCK, and CURE algorithms (Section 2.3.4). The assumptions mentioned above provide a rough estimation about what could be considered as high or low dimensionality and large or small data sizes. However, they remain to be assumptions and a lack of rules exist regarding these categorizations in academic literature. According to Dolnicar (2002) the only recommendation that could be given is to question if the dimensionality is not too high for the number of cases to be grouped (i.e., 2^k cases and preferably $5*2^k$). Table 1 provides an overview of the clustering algorithms with respect to the data characteristics as described in section 2.3.4.

Section 2.3.5 proposed a two-stage clustering approach to determine the number of clusters and obtaining accurate clustering solutions. The Framework in Table 2 outlines various strategies for User Profiling based on unsupervised machine learning and the data properties

including the data type, data size, and dimensionality. The framework includes strategies for categorical, numerical, and mixed types of data. The first stage consists of an hierarchical or model-based clustering procedure to determine the number of clusters and identify initial seeds. Secondly, a non-hierarchical clustering procedure is applied to provide more accurate cluster memberships.

Table 1

Overview of clustering algorithms and data characteristics as reviewed in section 2.3.4

Category	Algorithm	Data Type	Data Size	Handling High Dimensionality	Handling Noise
Model-Based Algorithms	SOMs (Kohonen, 1998)	Multivariate Data	Small/Moderate	Yes	No
Hierarchical Algorithms	Chameleon (Karypis et al., 1998)	Categorical/Numerical	Large	Yes	No
	ROCK (Guha et al., 2000)	Categorical/Numerical	Large	No	No
	CURE (Guha et al., 1998)	Numerical	Large	Yes	Yes
	Complete Linkage/Ward's (Tamasauskas et al., 2012; Pandove et al., 2018;)	Dependent on Distance Measure	Small/Moderate	No	No
Non-Hierarchical Algorithms	K-modes (Huang, 1998)	Categorical	Large	Yes	No
	K-medoids (Park et al., 2009)	Categorical	Small	Yes	Yes
	K-means (MacQueen, 1967)	Numerical	Large	No	No
	K-prototypes (Huang, 1998)	Categorical/Numerical	Large	Yes	No

Note. Adapted from Fahad et al. (2014)

Table 2

Framework outlining various strategies for User Profiling based on Two-Stage clustering and the characteristics of the dataset

Data Type	Data Size	Dimensionality	Stage - 1	Stage - 2
Categorical	Large	High	Chameleon	K-modes
		Low	ROCK	K-modes
	Small/Moderate	High	Chameleon	K-modes/K-medoids
		Low	Complete Linkage/Ward's	K-modes/K-medoids
Numerical	Large	High	CURE	K-means
		Low	CURE	K-means
	Small/Moderate	High	SOMs	K-means
		Low	SOMs	K-means
Categorical/Numerical (Mixed)	Large/Small	High	Chameleon	K-prototypes
		Low	ROCK	K-prototypes

Note. A data size of <=4000 is considered to be moderate/small. High Dimensionality is approximately >10 variables and Low Dimensionality is <=10 variables. A lack of rules exists regarding these data properties in literature (see section 2.3).

2.7 A Multi-Criteria Evaluation Model for User Profiling

The majority of literature focuses on the development of methodologies to data-driven segmentation and different types of user profiling approaches (e.g., Dynamic or Static). However, the requirements for effective user profiling have received less attention. As discussed in section 2.4, easily understandable profiling approaches generate ineffective segments while the complex approaches are accurate but not easy to use in practice (Cunfoglu, 2014). Additionally, the user profiling methods fail to acknowledge how different types of user data (i.e., implicit and explicit) affect the managerial usefulness of a segmentation base and User Profiling approach. For instance, Dolnicar (2009) studied the fundamental understanding of data-driven segmentation methodologies among marketing managers. The study concluded that 65 per cent of the study subjects had difficulties in understanding data-driven segmentation solutions and lacked a basic understanding about data-driven segmentation methodologies. More recently, Boratto et al. (2016) argued that the understand-ability or interpretability of data-driven segmentation methods continued to be an important research gap due to a lack of guidance by literature and the increasing complexity of the segmentation bases (e.g., amount of attributes). This issue can be referred to as the *managerial usefulness* of a segmentation base or user profiling approach. The managerial usefulness of segmentation results are higher when it is easy to interpret while the quality or accuracy of the profiles is high.

The first criteria for effective segmentation approaches were introduced by Thomas (1998) who argued to consider measurability, accessibility, stability, and sustainability (as cited in Goyat, 2011). However, more recent studies omitted measurability but have add the principle of attractiveness by including identify-ability, action-ability and responsiveness (van der Zanden et al., 2014; Wedel & Karmakura, 2012; Jadczakova, 2013). The evaluation criteria for effective user profiling are operationalized as follows: *Identify-ability* refers to approaches and attributes whereby profiles can be easily distinguished from each other on the basis of information that is obtained objectively (Zanden et al., 2014). Segments should be recognized easily so that they can be measured.

Substantiality is satisfied when the segments or profiles represent a large enough portion of the market to ensure profitability of behavioural targeting and other marketing efforts. This is related to the marketing goals and cost structure of an organization. Personalization becomes more prevalent due to advancements in information technology. To its limit, substantiality can be applied to each individual (or profile) where the purpose is to target each individual who produces revenues greater than the costs of the firm.

Accessibility is the degree to which users (i.e., marketing managers) are able to address the targeted segments with marketing efforts based on the customer attributes. Accessibility depends greatly on the availability and accuracy of (secondary) data sources and types of user data to generate user profiles or segments according to specific customer attributes (van der Zanden et al., 2014; Wedel & Karmakura, 2012)

Responsiveness is when users within the profile respond uniquely different from other profiles to marketing efforts. Hence, this is an important aspect for the effectiveness of any user profiling approach because differentiation in marketing efforts are more effective when each user profile is homogeneous in terms of customer behaviour and thus uniquely responds (van der Zanden et al., 2014; Wedel & Karmakura, 2012).

Stability of the user profiles or customer attributes is necessary for a long enough period in order to discover the user profiles, implement marketing strategies, and produce results. When the profiles to which a marketing effort is targeted change their behaviour or interests during the implementation, it is more likely not to succeed (van der Zanden et al., 2014; Wedel & Karmakura, 2012)

Action-ability is satisfied when the identified user profiles provide guidance for strategic decisions on effective marketing strategies. It differs from identify-ability or responsiveness

which only states that segments should be recognized easily and respond uniquely whereas the focus of action-ability is whether the segments (e.g., interests or needs) are consistent with the goals and core competencies of the organization (e.g., offerings) (van der Zanden et al., 2014; Wedel & Karmakura, 2012).

As discussed in section 2.5.1, customer segmentation variables can be classified into four major areas of geographic, demographic, psychographic, and behavioural variables (Goyat, 2011; Chen & Stallaert, 2014; Tsiptsis & Chorianopoulos, 2011; Hiziroglu, 2013). However, some researchers classify the segmentation base according to the level to which the variables can be observed. A segmentation base can be defined as a set of variables or characteristics used to assign potential customers to homogenous groups (Wedel & Karmakura, 2012). According to Jadczakova (2013) segmentation bases can be broadly classified into *observable* (i.e., measured directly) and *unobservable* customer attributes which can include either general or product-specific features. The most frequently used are observable customer attributes which are often referred to as *geo-demographics*. An advantage of this base is that such data can be easily collected and identified (Wedel & Karmakura, 2012). Geo-demographics provide segments that are considerably stable (e.g., gender), substantial (e.g., education, country), easy identifiable by objective measures (e.g., age), and easy to understand and interpret (van der Zanden et al., 2014). Furthermore, geo-demographics are readily available and can be used as a basis for User Profiles which provide managers with information on the accessibility of consumers (Jadczakova, 2013; van der Zanden et al., 2014). However according to Jadczakova (2013) a drawback of geo-demographics is their low responsiveness and action-ability due to clustering regions or neighbourhoods instead of individual customers. Next, *psychographics* can be classified as unobservable customer attributes (Hiziroglu, 2013). Psychographics aim to capture a customer's psyche, values, lifestyle, perceptions, and personality traits (Jadczakova, 2013). Psychographics form lifelike descriptions of consumers which enables marketers to translate a customer's triggers into marketing actions (i.e., very good action-ability and responsiveness) (Wedel & Karmakura, 2012). However, the stability of these segments are moderate and the accessibility is poor (Jadczakova, 2013; van der Zanden. 2014). Lastly, *behavioural variables* indicate the actual customer behaviour and interaction towards products and services including their needs and interests. These variables are classified into the unobservable segmentation base. Customers can be divided according to their identified behavioural and usage patterns. This type of segmentation is typically used to develop personalized offerings (Tsiptsis & Chorianopoulos, 2011). Examples are benefits sought, interests, preferences, intentions, brand loyalty, user status, or readiness to buy (Goyat, 2011; Jadczakova, 2013). Similar to Psychographics, behavioural-based variables develop highly responsive and action-able segments since they demonstrate significant differences in attitudes, needs, or interests which enables managers to develop tailored marketing campaigns. However their accessibility is limited because of weak associations with more general customer attributes such as geo-demographics (Hilziroglu, 2013; Jadczakova, 2013). Furthermore, the stability and identifiability are moderate since these variables are affected by the dynamic and implicit nature of a customers' needs and interests (Wedel & Karmakura, 2012). Hence, using variables of multiple segmentation bases can provide a more robust picture of the segments or profiles through which marketers can develop more effective and efficient marketing campaigns.

In brief, there are six evaluation criteria for effective customer segmentation. Moreover, a segmentation base can be broadly divided into observable variables and unobservable variables. Various User Profiling approaches are available including are Explicit Profiling, Implicit Profiling, and Hybrid Profiling. The majority of research focused on developing user profiling models and data-driven segmentation methodologies, but none considered to consider the requirements for effective user profiling. Furthermore, the understandability and interpretability of data-driven segments is more difficult due to an increasingly complex

segmentation base and a lack of guidance by literature. The proposed model in Figure 2 contributes towards these problems to some degree as it enables for a multi-criteria evaluation on different segmentation bases and types of user profiling based on implicit and explicit user data. The model is aimed at supporting both researchers and practitioners to determine which segmentation approach is appropriate for developing high quality profiles that are easy to interpret and utilize for marketing purposes. The model includes the six criteria discussed above which are considered to be essential for effective User Profiling. For instance, this paper segments website visitors according to ten behavioural attributes that indicate a visitors interests (i.e., implicit user data). From the model can be observed that these attributes yield profound insights which result in profiles that are highly actionable and responsive. However, the resulting profiles are less easy to understand, less stable, and less accessible. Combined with geo-demographics (e.g., country or study programme) makes the profiles easier to identify, understand, to access, and more actionable. An overview of the segmentation bases are depicted in Table 3 and the model is depicted in Figure 2. Nevertheless, the appropriateness of any User Profiling approach and segmentation base is highly dependent on the knowledge discovery goal.

Table 3
Overview and Example of Possible Segmentation Bases and Customer Attributes

Segmentation Base	General Attributes	Specific Attributes
Observable	Geo-Demographics, Culture, Socio-Economics	User Feedback, Usage Frequency, Loyalty, Readiness to Buy
Unobservable	Personality Traits, Values, Lifestyle	Intentions, Preferences, Interests, Perceptions, Benefits Sought

Figure 2. A Multi-Criteria Evaluation Model for User Profiling

Segmentation Base	Type of User Profiling	Identifi-ability	Substan-tiality	Stability	Access-ibility	Respon-siveness	Action-ability	Understand-ability
Demographic and Geographic (Observable)	Explicit User Profiling	Very Good	Very Good	Good	Good	Poor	Poor	Good
Psychographics and Behaviour-Based (Unobservable)	Implicit User Profiling	Moderate	Good	Moderate	Poor	Very Good	Very Good	Poor
Mixed Segmentation Base	Hybrid User Profiling	Very Good	Good	Moderate	Good	Very Good	Very Good	Moderate

3. METHODOLOGY

This study aims to discover online behavioural profiles among Dutch website visitors interested in Master studies at the UT. The process of Knowledge Discovery in Databases (KDD) used as a research methodology. The process includes the following iterative steps: Understanding the application domain, Select Target Data, Data Pre-processing, Data Transformation, Data Mining, Interpretation and Evaluation, and Consolidating the Discovered Knowledge. The methodology chapter includes the first 5 steps of the KDD process. Chapter 4 and 5 consist of the remaining two steps. The proposed framework and model are used in order to obtain the most reliable results and realizing the goal of the paper.

3.1 Understanding the Application Domain

The UT aims to educate the professionals of the future. The educational institute distinguishes itself from other institutes by offering both technical and social studies. In total 10.435 students enrolled in 2017. Approximately 5.489 are Bachelor applicants and 4.010 Master applicants (Facts & Figures, 2018). The UT includes about 79 nationalities whereby most are Dutch (Facts & Figures, 2018). In total, there are five faculties, 20 bachelor programmes, and about 33 Master programmes (Facts & Figures, 2018). The M&C department of the UT is among others responsible for monitoring the Higher Education market (HE) developments and developing (online) student recruitment campaigns. A lot of data is collected and stored in the UT back-end systems and the activities a lead is taking on the UT website are tracked. *Osiris Application Submitted* is an important conversion point as it indicates whether a visitor fully completed the application process. Until now the M&C department was not able to find the right structure in the data to discover behavioural profiles. In addition, the HE market has to cope with increasing competition to recruit students. Marketing concepts which have been effective in business, are now needed by many universities looking to gain a competitive edge and gaining market share (Hemsley-Brown, & Oplatka, 2006). Changes in the HE market are, among others, caused by the increasing cost of education, globalization, or numerus fixus (Barber, Donnelly, Rizvi, & Summers, 2013). Furthermore, Barber et al. (2013) argues that it is of increasing importance that "each university needs to be clear which market segments it wants to serve and how" (p. 5). Additionally, potential applicants face complex challenges of narrowing down personal interests into a single HE programme. Therefore, it is important for the M&C department to benefit from BT and ML techniques in order to be more efficient and effective in their targeted marketing efforts. The outcomes would ideally unveil high converting behavioural profiles among Dutch website visitors of the University of Twente interested in Master studies.

3.2 Target Data and Pre-Processing

There are about 79 nationalities among UT students which make it nearly impossible to analyse all of them in a limited timeframe. Therefore, this study selects a country (i.e., A priori) where the majority of the traffic of prospective students arrive from at the UT website. Dutch students are the largest group of applicants for the UT and belong to the majority of the website visitors (Facts & Figures, 2018). Furthermore, the UT has to cope with increasing competition in the (Dutch) HE market and would like to tailor their marketing efforts to effectively and efficiently reach relevant potential prospects by data-driven insights. In order to yield the most accurate profiles the researcher further specifies the target data to only include visitors interested in Master studies. To discover important differences among behavioural profiles the selected data consists of two groups: (1) all website visitors interested in Master studies and (2) all Dutch website visitors interested in Master studies. These groups are referred to as *All Master Visitors* and *Dutch Master Visitors* from here on.

The *selected target data* is of secondary nature meaning that the data was not collected first hand but extracted from the UT's back-end systems. The data consists out of Excel databases wherein the rows represent website visitors and the columns represent the behavioural

attributes. Between *October 2016* and *August 2017, 32.495* observations have been collected by the *CRM-system*. Filtering the database to only include visitors interested in UT Master studies, removing duplicates, and missing values resulted in 5962 unique observations. Next, the CRM data is *combined* with the data extracted from Google Analytics. Within the same period, *57.598* observations were extracted with *Google Analytics* using the GA add-on to extract *3-4 weeks* of data per report, and ensuring a *sample percentage* between 83.3% and 100%. Website visitors are automatically anonymized by a unique user ID (wrd-id) assigned by the UT back-end systems. The unique ID's are used to combine the CRM and Web data in Excel with matching wrd-id's resulting in a clean dataset of *3612 observations* for All master visitors and *412 observations* for Dutch Master visitors. Lastly, this study adopts the hybrid approach of user profiling. The data extracted from Google Analytics is an implicit form of user data such as PDF downloads. Explicit user data is extracted from the UT's CRM system such as the type of study programme.

3.3 Data Transformation

The pre-processing of the raw data sources generated many attributes that are potentially useful to discover behavioural profiles among Dutch website visitors interested in Master studies. The Osiris Application Submitted is an important conversion point as it indicates whether a visitor fully completed the application process. All extracted attributes are divided into five main categories namely, *(onsite) behavioural attributes, traffic source, country, study programme,* and *preferred device type*. The first category is the type of online behaviours manifested by visitors of the UT website. The traffic source category indicates how visitors found their way to the UT website (e.g., referrals). The third category includes nearly two hundred different countries of the UT website visitors. The fourth category consists out of approximately thirty-one Master studies wherein visitors are interested. Finally, the fifth category includes the device type used to surf the UT website. Ten behavioural attributes have been extracted from the pre-processed data and are listed below. The remaining categories are used to describe the discovered profiles in more detail. Furthermore, the data consists of categorical attributes which are transformed into binary (0-1) attributes. A binary value of 1 indicates a presence of a particular behaviour and 0 indicates an absence. In this study a symmetric binary dataset is analysed including the following attributes:

Osiris Application Submitted	Managed CTA click
Educational Brochure Request	PDF Download
Eligibility Check	Scholarship Finder
Questions via Web Form	Open Day Registration
Request Student for a Day	Frequently Asked Questions

Osiris Application Submitted indicates whether a visitors fully completed the application process via the student information system Osiris. *Educational Brochure Request* is triggered when website visitors downloaded at least one kind of educational brochure. The *Eligibility Check* informs visitors whether they meet the minimum requirements for a particular UT study. Furthermore, it allows the UT to obtain more detailed information and to identify common strengths or weaknesses from its prospects. *Questions via Web Form* is manifested when a visitor asked a question via web forms on the UT webpages. The *Request Student for a Day* attribute implies that a visitor registered to try out or experience a day at the UT with a current UT student as a guide. *Managed CTA Click* is a call-to-action attribute to motivate a visitor to take a certain action. In this study, a Managed CTA is when a visitor revisits the UT website whereby the content or message of the CTA are automatically tailored to motivate the visitor to take action (e.g., Register Now). *PDF download* is triggered when a visitor downloads any form of PDF that is non-study related. Examples are financial information, a ground plan,

particular events, or a catalogue of the UT. The *Scholarship Finder* allows visitors of the UT website to search for available scholarships. An *Open Day Registration* is when a visitor registered for an open day at the UT. Finally, the *Frequently Asked Questions* (FAQs) attribute is manifested when a visitor considered the FAQ page for information.

3.4 Data Mining

Following the proposed framework, the dataset is considered as being small to moderate sized including 3612 observations with a low dimensionality of 10 behavioural attributes. Therefore, hierarchical clustering using complete linkage followed by the non-hierarchical algorithm k-modes is adopted. The hamming distance is used as a similarity measure for the symmetric binary dataset. First, analysis is conducted on All Master visitors to obtain a comprehensive view of the data structure and behavioural profiles. Secondly, in order to discover behavioural profiles of Dutch Master visitors the variable country is used and analysed by the same two-stage clustering methodology. Figure 3 illustrates the steps taken for the first analysis of All Master Visitors and Figure 4 illustrates the steps taken for the second analysis of Dutch Master Visitors. *R statistical programming* is used for data mining and *Microsoft Excel* is used for visualization of the results. R provides more freedom for analysis as opposed to click-supported programmes and allows to tailor an algorithm to the specific goals of this research. The detailed process is as follows: first, the Hamming Distance is calculated with the 'hammingD' function of the package 'EnsCat'. Secondly, the 'stats' package and 'hclust' function are used for the hierarchical clustering procedure. Thirdly, the sum of within-cluster inertia is calculated with the 'best.cutree' function of the package 'JLutils' to support the determination of an appropriate number of clusters. In stage 2, the k-modes algorithm is applied by inserting the number of clusters as determined in stage 1 with the package 'KlaR' and function 'kmodes()'. The results are then written to the dataset with the function write.xlsx() of the package 'openxlsx'. An overview of the complete process described in this chapter for conducting the analysis is depicted in Figure 5.

Figure 3. Steps taken for analysis of All Master Visitors

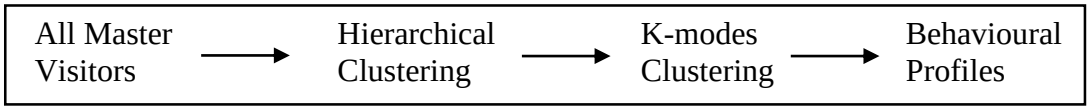


Figure 4. Steps taken for analysis of Dutch Master Visitors

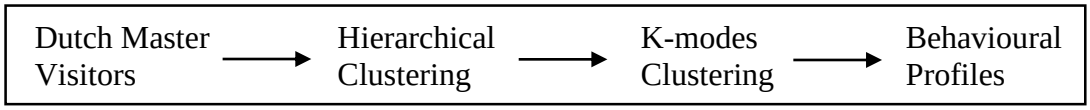
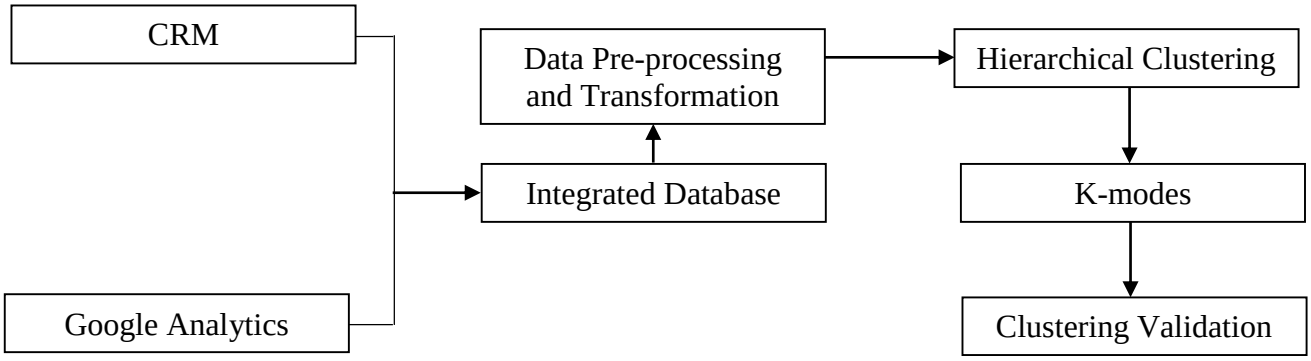


Figure 5. Abstract illustration of the process for conducting the analysis



3.5 Data Protection Regulations

Privacy is an important issue, for in this case, the field of (data-driven) marketing. Previous studies already stated that visitors are not sufficiently aware about the extent and possibilities of online tracking and targeting technologies which threatens their privacy (e.g., Goldfarb & Tucker, 2011; Borgesius, 2016). For instance, Facebook and Cambridge Analytica (i.e., A British political consulting firm) used personal data of millions of users without their consent for political purposes on Donald Trump's US presidential election campaign and to influence the Brexit vote in 2016. This moment significantly impacted the public understanding and awareness of personal data. Since May 25th 2018 the General Data Protection Regulation (GDPR) was enforced (European Commission, 2018). The GDPR is a European Legislation on the protection of personal data and the free movement of such data. The data in this study is carefully handled in compliance with the GDPR to avoid breaching the regulations and privacy of UT website visitors. The data was anonymized by the UT back-end systems by assigning each new visitor a unique user ID (WRD-ID). Moreover, the University of Twente announced have informed its subjects about how and where the data will be used as well as providing guidance on the protection of personal data in scientific research. To the best of the authors knowledge, this study meets the requirements of the General Data Protection Regulation.

3.6 Cluster Validation

An important aspect of clustering analysis is to validate the clustering results. Moreover, the data is of secondary nature and the reliability and quality of such data must be critically assessed. A drawback of clustering analysis is that the algorithms always generate clusters but their outcomes might not always accurately reflect the goals of the research or analysis. This is important as it indicates whether the visitors of the UT website and their behaviour are accurately grouped. Various tests are available in literature for validating the clustering solutions but none of them performed better than the other (Arbelaitz et al., 2013). However, two of the most commonly used tests are used in this study namely, the *Silhouette's test* (i.e., homogeneity test) and *cross-validation*. There are three fundamental concepts of clustering validity namely, external criteria, internal criteria, and relative criteria (Halkidi et al., 2001). External criteria evaluates the results of a clustering solution based on a pre-specified structure that is projected on the dataset and reflects the user's intuition about the structure in the data (Halkidi et al., 2001). Internal criteria evaluates the quality of the clustering structure in terms of quantities that involve vectors of the dataset such as in a proximity matrix (Halkidi et al., 2001). Relative criteria evaluates the clustering structure by comparing it with the same algorithm but different parameter values (e.g., compactness and separation).

The *silhouette score* is calculated by measuring how closely each cluster member is located to its profile centroid (Amorim & Henning, 2015). Furthermore, it measures the average distance between clusters and the degree to which the observations are well structured (Jain, 2016; Amorim & Hamming 2015). A score of 1 implies that the cohesion within the clusters is quite good. A score closer to -1 indicates that the cohesion within the clusters is poor and not as valid. Hence, the score indicates how well visitors are distributed within clusters (Amorim & Hamming, 2015).

Cross-validation can be done according to three variations namely, *hold out, K-folds, and Leave-one-out* (Schneider, 1997). This paper adopts the hold out method. However, it is important to note that the hold-out method is a simple variation of cross-validation. It involves only a single run whereas more exhaustive methods run several times on multiple *k*-partitions. Hence, the hold-out method may include some variation depending on how the data is randomly split. Cross-validation will be performed for All Master Visitors and Dutch Master Visitors by randomly splitting the original data into two sub-samples. The sub-samples consist of a training dataset (75%) and a test dataset (25%). The hold out variation of cross-validation is used to evaluate the appropriateness of the chosen number of clusters.

4. RESULTS

4.1 Descriptive Statistics

The pre-processed dataset consists out of 3612 UT website visitors interested in Master studies and 10 behavioural attributes. In Table 4, the range of the behavioural attributes equals to one, because all categorical attributes are transformed into binary (0-1) attributes. A binary value of 1 indicates a presence of a behaviour and 0 indicates an absence. Therefore, each attribute in Table 4 has a minimum value of zero and a maximum value of one. Considering the nature of binary attributes the minimum and maximum might not be very meaningful. However, it provides an indication if any error exists in the dataset. Table 4 indicates no irregularity or error in the dataset when evaluating the range, minimum, and maximum.

Table 4

Descriptive Statistics of pre-processed database

Behavioural Attributes	N	Range	Minimum	Maximum	Mean
Educational Brochure Request	3612	1	0	1	.41
Eligibility Check	3612	1	0	1	.69
Questions via Web form	3612	1	0	1	.12
Request Student for a Day	3612	1	0	1	.02
Osiris Application Submitted	3612	1	0	1	.05
Managed CTA Click	3612	1	0	1	.10
PDF Download	3612	1	0	1	.05
Scholarship Finder	3612	1	0	1	.05
Open Day Registration	3612	1	0	1	.02
FAQs	3612	1	0	1	.02
Valid (N)	3612				

The behavioural attributes in Table 4 can be described in more detail by the factors country, study programme, device type, and traffic source. The means in Table 4 refer to the proportion of each behavioural attribute in the data. For example, approximately 69% of the website visitors took the eligibility check. 41% of the visitors requested an educational brochure, 12% asked a question via web form, and about 10% clicked on a managed CTA. Furthermore, approximately 5% submitted their application in Osiris. The lowest means in Table 4 are 2% which translates to about 72 visitors who conducted a particular behaviour, such as an open day registration. There are no rules-of-thump about the sample size necessary for clustering analysis (Dolnicar, 2002). However, one approach to determine the minimum sample size is to include no less than 2^k cases (k = number of variables), and preferably $5*2^k$ (Dolnicar, 2002). In this study the number of behavioural attributes is 10. Therefore, the sample size in this study should be at least between 1024 and 5120 according to the suggested method.

The dataset in this study includes 3612 observations and meets the requirements of conducting clustering analysis. No irregularities or outlies are present in the dataset. Furthermore, standardization of the data is not required because the categorical attributes are transformed into binary attributes where the range, minimum, and maximum of all behavioural attributes are identical. As a result, the dataset is ready for analysis. The analysis in this study is two-fold: first, hierarchical clustering is used to determine K the number of clusters and identify initial seeds. Secondly, the non-hierarchical clustering algorithm k-modes is applied to provide more accurate cluster memberships.

4.2 Determining the Number of Clusters

Hierarchical clustering is applied in order to determine an appropriate number of clusters. In this paper, a symmetric binary dataset is analysed. For this type of data, Tamasauskas et al. (2012) evaluated the performance of ten different hierarchical clustering methods by experimenting with ten different distance measures in terms of accuracy. As discussed in chapter 2, complete linkage in combination with the hamming distance performed best on symmetric binary data (Tamasauskas et al., 2012). Hence, this study adopts the aforementioned combination. First, the hamming distance is calculated with the 'hammingD' function of the package 'EnsCat'. Secondly, the 'stats' package and 'hclust' function are used for the hierarchical clustering procedure. Lastly, the sum of within-cluster inertia is calculated with the 'best.cutree' function of the package 'JLutils' to further support the determination of the number of clusters to be selected.

Figure 6 and 7 present the number of clusters based on the dendrograms and relative loss of inertia method by calculating the sum of within-cluster inertia for each partition (Husson et al., 2018). The best partition is accentuated in black and the second-best in grey. Figure 6 indicates that for the best partition the tree should be cut into 4 clusters and for the second-best into 3 clusters for All Master visitors. However, the dendrogram in figure 6 depicts that 6 clusters is appropriate. The difference between the relative loss of inertia between 4 or 6 clusters is nearly 0.0007 and negligible. Figure 7 illustrates that for Dutch Master visitors the best partition to cut the tree is 3 clusters and the second-best is 6 clusters. The dendrogram of Dutch Master visitors illustrates that 6 clusters is appropriate. Hence, it can be concluded that 6 clusters is appropriate for both groups. The outcomes of the non-hierarchical clustering procedure are presented in the following section. The number of clusters and clustering results are validated in chapter 4.5.

Figure 6. Dendrogram and relative inertia loss of All Master Visitors

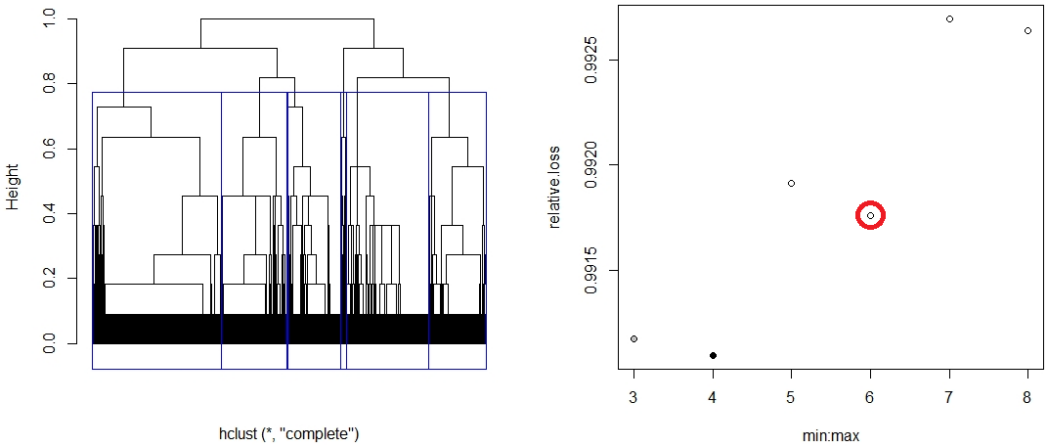
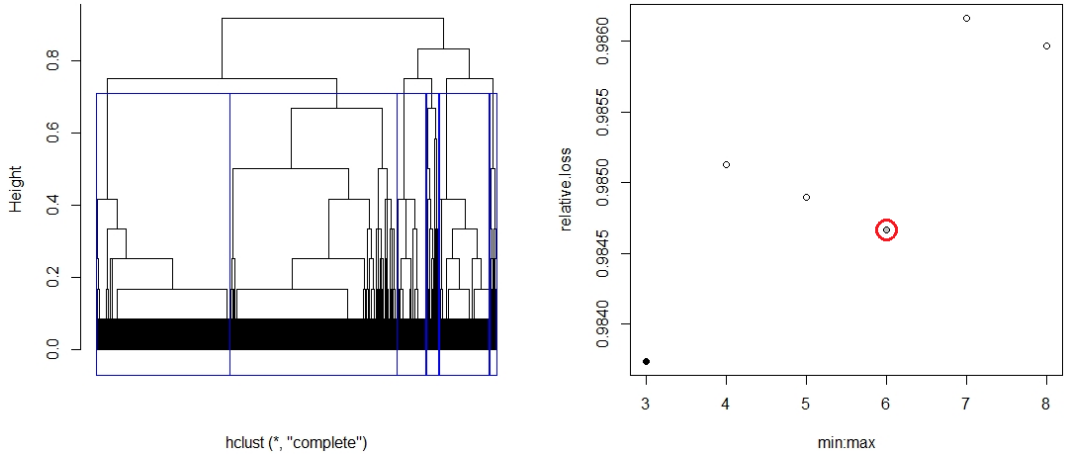


Figure 7. Dendrogram and relative inertia loss of Dutch Master Visitors



4.3 Cluster Analysis

This chapter presents the evaluation phase of the data mining process and the behavioural profiles discovered after conducting the non-hierarchical clustering procedure. First, behavioural profiling is done on all website visitors interested in UT Master studies. Secondly, behavioural profiling is done on Dutch website visitors interested in UT Master studies to identify whether the discovered profiles are independent of country. This includes discovering possible similar and dissimilar characteristics between profiles. The results can provide valuable insights for the UT M&C department to increase the effectiveness and efficiency of marketing efforts. The explanatory variable is 'Osiris Application Submitted'. This is an important conversion point as it indicates whether a visitor fully completed the application process. Each individual visitor can manifest a different sequence of behaviour. For example, in a profile that includes high converting visitors, one can first request an educational brochure and then a PDF download or vice versa. As a result, the visitors in the profile show different sequences but share similar behaviours. When both visitors submitted their application, it is more meaningful to analyse the similarity between behavioural attributes than to analyse the specific sequence they followed.

The AIDA-model is used as a reference to recognize the stage of a visitor's customer journey, assign cluster labels, and describing each profile. The model consist of the Attention, Interest, Desire, and Action stage (Wijaya, 2012). In the *Attention* stage the visitors become aware of the service or product (i.e., UT studies) and seek to inform themselves by, in this case, information on the website. This stage is associated with cognition and rational knowledge seeking (Wijaya, 2012; Rawal, 2013). The *Interest* stage is associated with customers who like to acquire sufficient knowledge and have developed an affiliation for the institute to some degree (Wijaya, 2012; Rawal, 2013). For example, the Eligibility Check often demonstrates a visitor is interested as it informs them whether they meet the minimum requirements to be eligible for UT studies. In the *Desire* stage a visitor has developed a favourable attitude towards the institute. Examples of behaviours that relate to this stage are Open Day Registrations or Requesting to be Student for a Day. Furthermore, visitors in the Desire stage are naturally more engaged and thus more likely to manifest multiple behaviours prominently which are related to the previous stages. Different variations of the AIDA-model are available in terms of extensions to the original. However, the nature of the AIDA-model remains unchanged and many researchers and practitioners still adopt it today. Information technology enabled the emerge of various social media platforms that radically changed how customers socialize and communicate. Therefore, Waijaya (2012) the AISDALSLove-model. However, this study focuses on behavioural attributes that are manifested by visitors on the UT website and these attributes do not include social media attributes. Traffic Sources are included (e.g., Google, Facebook) but are regarded as referrals whereby actual social media usage or behaviour is not included. Therefore, the AIDA-model is an appropriate reference to recognize the stage of a visitor's customer journey, assign cluster labels, and describe each behavioural profile.

4.3.1 Behavioural Profiling of All Master Visitors

The non-hierarchical clustering method k-modes is applied by setting the number of clusters to 6 for All Master Visitors (N=3612) as determined by hierarchical clustering in section 4.2. Table 5 denotes that cluster 6 represents the majority of visitors by containing 49.5% of All Master Visitors. The second largest is cluster 4 with 23.4% followed by cluster 1 with 11.7%. This indicates that clusters 6, 4, and 1 contain about 85% of All Master Visitors. The remaining 15% is distributed in cluster 3 by 8.5%, cluster 2 by 3.8%, and cluster 5 by 3.1%. The smallest cluster (5) translates to 112 visitors and largest cluster (6) translates to 1789 visitors. It can be observed that there are no unassigned visitors to clusters.

Table 5
Distribution of All Master Visitors in each cluster

	N	%
Cluster 1	421	11.7
Cluster 2	137	3.8
Cluster 3	306	8.5
Cluster 4	847	23.4
Cluster 5	112	3.1
Cluster 6	1789	49.5
Total	3612	100.0

In the 1st cluster of Table 6, 14.0% converted by submitting their application (i.e., Osiris Application Submitted). Prominent attributes are 'Educational Brochure Request' (100%), 'Eligibility Check' (100%), and 'Questions via Web form' (15.4%) which implies that members of this cluster are interested in studying at the UT. Furthermore, 26.1% engaged with a managed CTA (i.e., Call-To-Action). Requesting Educational Brochures demonstrates a visitors interest to acquire more information which in turn raises their awareness about studying at the UT. Likewise, the Eligibility Check informs visitors whether they meet the minimum requirements for a specific study programme they are interested in. Furthermore, a reasonable amount of visitors asked a question via Web form (15.4%) which contributes to their awareness and supports the interpretation of them being interested. In contrast, asking questions can be indicative of missing information on the UT website. Cluster 1 manifests the second highest conversion score compared to all other clusters. Such characteristics resemble the interest phase of the AIDA-model. The phase is characterized by customers who like to acquire more information and at the same time they have developed an affiliation for the institute to some degree (Wijaya, 2012). Therefore, the members in cluster 1 can be labelled as '*Interested High Potential Prospects*' or '*Interested-HP*'. The majority of the members in cluster 1 are interested in Sustainable Energy Technology (15.9%), Mechanical Engineering (11.4%), and Civil Engineering and Management (11.2%) (Table 7). Furthermore, members in cluster 1 come from India (21.6%) followed by The Netherlands (10.5%), and Indonesia (5.0%) (Table 8). As depicted in Table 9 the traffic in cluster 1 comes from Google/Organic (53.2%), Quick link (19.5%), and Google/cpc (7.4%). Lastly, the most popular device type in cluster 1 is a Desktop (61.3%) followed by Mobile (17.2%) and Tablet (2.9%) (see Table 10).

In the 2nd cluster of Table 6, about 0.7% submitted their application. Prominent behavioural attributes in this cluster are 'Managed CTA click' (100%), 'Educational Brochure Request' (97.1%), and 'Scholarship finder' (41.6%) followed by 'Open Day Registration' (11.7%), and Request Student for a Day (2.9%). These characteristics resemble visitors with an interest to acquire information about UT studies, scholarships, and experiencing the student life at the UT. All members manifested a Managed CTA which implies that visitors came across UT advertisements and downloaded various forms of Educational Brochures for more detailed information. However, zero members in cluster 2 went through the Eligibility Check which can be indicative of visitors who require more information about the UT and study programmes before considering the Eligibility requirements of a particular study. The Scholarship Finder is the second manifested behaviour that characterizes cluster 2. It can be argued that Scholarship availability is an important factor when considering to study at the UT for this group of visitors. Therefore, it can be hypothesized that members of cluster 2 are interested in UT studies but in addition to the brochures their decision is dependent on scholarship availability. As this cluster is distinctive in terms of Scholarship oriented visitors compared to the other clusters, it can be labelled as *Scholarship-Driven*. Popular studies in cluster 2 of Table 7 are Health Sciences (12.4%), Spatial Engineering (10.2%), and Civil Engineering and Management (9.5%). The

majority of members in cluster 2 are from India (24.1%), The Netherlands (21.9%), and Nigeria (4.4%) (Table 8). Similar to cluster 1, the most popular device type in Table 10 is Desktop by 61.3%, followed by Mobile (39.4%) and Tablet (2.9%).

In the 3rd cluster, 1.6% converted on Osiris Application Submitted. Cluster 3 is mainly characterized by members who asked a Question via Web form (100%) followed by the Eligibility Check (67.3%). Zero members in cluster 3 have Requested an Educational Brochure. As all members asked a question via Web form it is highly probable that visitors did not obtain all the information they were searching for. A large proportion (67.3%) conducted the Eligibility Check which implies that visitors are interested to see whether they meet the minimum admission requirements for a study programme they are interested in. This behaviour resembles characteristics of visitors who are interested at studying at the UT to some degree but require more specific information that may not be available on the website. The UT M&C Department could apply text mining to analyse the kind of questions being asked and improve marketing efforts or website content. In cluster 3, no brochures were requested which decreases their awareness and all members asked questions via Web form to obtain the right information which implies their interest. Therefore, cluster 3 can be labelled as *Moderately Aware Potential Prospects* or *MA-P*. Popular studies in cluster 3 are Electrical Engineering (9.8%), Business Administration (9.2%), and Computer Science (9.5%). The majority is from India (16%), closely followed by the Netherlands (15%) and Germany (6%). Traffic mainly comes via Google/Organic (3.6%) and the remaining proportion of traffic sources are negligible. This implies a possible limitation of the veracity and volume of the dataset. Lastly, the majority of cluster 3 prefers to use a Desktop (3.6%).

In the 4th cluster, 0.8% submitted their application. It represents the second largest cluster in terms of size. All members requested an Educational Brochure (100%). Furthermore, (6%) used the scholarship finder, 5.8% downloaded some form of PDF, and 3.1% asked a question via Web form. However, zero members in cluster 4 have conducted the Eligibility Check and Managed CTA. The scores of the remaining behavioural attributes are negligible. A distinctive characteristic of this group is that all visitors requested an Educational Brochure which raises their awareness. Additionally, the majority of this group visits the website via Google/organic search (29.9%). It could be hypothesized that visitors became aware by a positive spread of Word of Mouth or advertisements and used Google with keywords that relate to the information on the UT website. Therefore, they have downloaded various forms of Educational Brochures. These characteristics resemble visitors that are in the Attention stage. In this stage customers become aware of the products and services and seek to acquire more information. Furthermore, this stage is associated with cognition and rational knowledge seeking (Wijaya, 2012; Rawal, 2013). Hence, this cluster is labelled the *Attention* group. Popular studies are Spatial Engineering (12.8%), Geo Information Science and Earth Observation (9.9%), and Environmental and Energy Management (9.8%). The majority comes from India (24.3%) and the Netherlands (17.2%). Lastly, the majority prefers a Desktop (37.0%).

In the 5th cluster, 78.6% converted on 'Osiris Application Submitted'. This is the highest conversion score compared to all other clusters. Dominant behavioural attributes are Eligibility Check (100%), Educational Brochure Request (91.1%), PDF download (77.7%), and managed CTA click (75.9%). Additionally, 36.6% used the scholarship finder, 25.9% asked a question via Web form, and 19.6% considered the FAQ page. Less prominent are Registration for Open Day and Request Student for a Day (2.7%). The majority of visitors are from India (19.6%). The latter could potentially explain why Registration for Open Days and Request Student for a Day are less prominent. For example, it might be too demanding in terms of financial expenses and visa arrangements to visit an open day or experience the student life for a day. Therefore, one could hypothesize that behavioural profiles of high converting visitors might differ between national and international students in terms of Open Day Registration and Requesting Student

for a Day. At least 8 of 10 visitors manifested 5 out of 10 behaviours which demonstrates their high level of interest compared to other clusters. Approximately 3 out of 10 either considered scholarship availability, asked a question via Web form, or visited the FAQ page to obtain more detailed information. This behaviour is distinctive for their degree of interest and desire to study at the UT. The high conversion percentage can potentially be explained by the fact that about 8 out of 10 touchpoints have been conducted. These characteristics resemble the Desire and Action stage. The desire stage is characterized by the development of a favourable attitude towards the institute (i.e., affection) and the action stage by the intention to perform a behaviour or the behaviour itself (i.e., conation). As the purpose is to motivate all relevant visitors to submit their application, this cluster can be labelled as the *Desired High Potential Prospects* or *Desired-HP*. Popular studies are Civil Engineering and Management (18.8%), Sustainable Energy Technology (17.0%), and Mechanical Engineering (15.2%). The majority comes from India (19.6%) and the Netherlands (16.1%). Most traffic comes through Google/Organic (81.3%), E-mail (19.6%), and Quick Link (18.8%). Cluster 5 is the only cluster where E-mail and Quick Link are more prominent after Google/Organic compared to other clusters. This behaviour potentially explains the higher Managed CTA percentage. Lastly, the preferred device type is a Desktop (83.9%).

In the 6th cluster, 0.5% submitted their application. Cluster 6 is the largest cluster in terms of size. The cluster is characterized by the Eligibility Check that have been conducted by all cluster members (100%). As the vast majority only performed one behaviour it can potentially explain why this cluster has the lowest conversion rate. Registration for Open Days have been conducted by 0.6% and Request Student for a Day by 0.4%. The remaining behavioural attributes are negligible. Conducting the Eligibility Check suggests that members of this cluster are interested to see whether they meet the minimum requirements for the study programme they are interested in. There is an extremely small proportion of visitors that conduct any kind of other behaviour to acquire information about the UT. Therefore, it is highly probable that visitors in cluster 6 obtained information about the UT from sources other than the website. Potential sources that can raise the awareness and interest for studying at the UT are a positive spread of Word of Mouth (WOM), Electronic Word of Mouth (eWOM) through social media channels, and online and offline advertising. The majority of the members in cluster 6 are from India (18.8%) and the Netherlands (7.2%) (Table 8). Interestingly, the difference between the two countries seems to be substantially larger in comparison to other clusters. In other clusters the majority also consists of Indian visitors followed by the Netherlands, except, the differences are smaller. Hence, Indian visitors in cluster 6 could be more likely to raise their awareness using external sources and are mainly interested if they meet the minimum requirements to be eligible compared to visitors of other nationalities. These characteristics resemble the interest stage of the AIDA-model which is associated with customers who have developed an affiliation for the institute to some degree (Wijaya, 2012). Hence, cluster 6 is labelled as *Interested* prospects. Popular studies are Mechanical Engineering (10.9%), Computer Science (9.5%), and Electrical Engineering (9.5%). The preferred device type is a desktop (11.2%) and the major traffic sources is through Google.

Table 6

Distribution of Behavioural Attributes of All Master visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Interested-HP		Scholarship-Driven		MA-P		Attention		Desired-HP		Interested	
Behavioural Attributes	N	%	N	%	N	%	N	%	N	%	N	%
Educational Brochure Request	421	100.0	133	97.1	0	0.0	847	100.0	102	91.1	0	0.0
Eligibility Check	421	100.0	0	0.0	206	67.3	0	0.0	112	100.0	1789	100.0
Questions via Web form	65	15.4	7	5.1	306	100.0	26	3.1	29	25.9	0	0.0
Request Student for a Day	11	2.6	4	2.9	5	1.6	6	0.7	3	2.7	8	0.4
Osiris Application Submitted	59	14.0	1	0.7	5	1.6	7	0.8	88	78.6	9	0.5
Managed CTA Click	110	26.1	137	100.0	6	2.0	0	0.0	85	75.9	8	0.4
PDF Download	26	6.2	13	9.5	3	1.0	49	5.8	87	77.7	2	0.1
Scholarship Finder	42	10.0	57	41.6	1	0.3	51	6.0	41	36.6	1	0.1
Open Day Registration	13	3.1	16	11.7	8	2.6	4	0.5	13	11.6	10	0.6
FAQs	14	3.3	5	3.6	1	0.3	7	0.8	22	19.6	1	0.1

Table 7

Top Three Study Programmes of All Master Visitors in each cluster

Cluster 1 Interested-HP			Cluster 2 Scholarship-Driven			Cluster 3 MA-P			Cluster 4 Attention			Cluster 5 Desired-HP			Cluster 6 Interested		
Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%
SET	67	15.9	HS	17	12.4	EE	30	9.8	SE	108	12.8	CEM	21	18.8	ME	195	10.9
ME	48	11.4	SE	14	10.2	BA	28	9.2	GISEO	84	9.9	SET	19	17.0	CPS	170	9.5
CEM	47	11.2	CEM	13	9.5	CPS	27	8.8	EEM	83	9.8	ME	17	15.2	EE	170	9.5

Note. Explanations of abbreviations are available on p.3. Complete list of studies per cluster are available in Appendix H.

Table 8

Distribution of Top 3 Visitor Countries of All Master Visitors in each cluster

Cluster 1 Interested-HP			Cluster 2 Scholarship-Driven			Cluster 3 MA-P			Cluster 4 Attention			Cluster 5 Desired-HP			Cluster 6 Interested		
Country	N	%	Country	N	%	Country	N	%	Country	N	%	Country	N	%	Country	N	%
Europe																	
Netherlands	44	10.5	Netherlands	30	21.9	Netherlands	46	15.0	Netherlands	146	17.2	Netherlands	18	16.1	Netherlands	128	7.2
Germany	18	4.3	United Kingdom	2	1.5	Germany	19	6.2	Germany	26	3.1	Spain	4	3.6	Germany	73	4.1
Italy	11	2.6	Spain	2	1.5	United Kingdom	13	4.2	Italy	10	1.2	Greece	4	3.6	Greece	39	2.2
Asia																	
India	91	21.6	India	33	24.1	India	49	16.0	India	206	24.3	India	22	19.6	India	336	18.8
Indonesia	21	5.0	Pakistan	7	5.1	Pakistan	15	4.9	Indonesia	53	6.3	Iran	5	4.5	Pakistan	101	5.6
Pakistan	15	3.6	Indonesia	7	5.1	Iran	15	4.9	Bangladesh	13	1.5	Indonesia	5	4.5	Indonesia	91	5.1
North America																	
United States	14	3.3	Mexico	2	1.5	United States	16	5.2	United States	14	1.7	Mexico	5	4.5	United States	64	3.6
Mexico	13	3.1	Honduras	1	0.7	Mexico	4	1.3	Mexico	5	0.6	United States	3	2.7	Mexico	28	1.6
Costa Rica	3	0.7	Canada	1	0.7	Canada	4	1.3	Canada	4	0.5	Canada	1	0.9	Canada	18	1.0
South America																	
Brazil	9	2.1	Brazil	3	2.2	Brazil	6	2.0	Colombia	16	1.9	Brazil	3	2.7	Brazil	36	2.0
Colombia	5	1.2	Paraguay	1	0.7	Colombia	2	0.7	Brazil	8	0.9	Venezuela	1	0.9	Colombia	13	0.7
Suriname	2	0.5	Ecuador	1	0.7	Venezuela	1	0.3	Ecuador	3	0.4	Peru	1	0.9	Venezuela	6	0.3
Africa																	
Nigeria	14	3.3	Nigeria	6	4.4	Nigeria	7	2.3	Nigeria	38	4.5	Nigeria	4	3.6	Nigeria	70	3.9

Ghana	13	3.1	Ghana	3	2.2	South Africa	5	1.6	Ethiopia	17	2.0	Ghana	4	3.6	Ghana	60	3.4
South Africa	7	1.7	South Africa	2	1.5	Ghana	3	1.0	Kenya	16	1.9	Kenya	3	2.7	Ethiopia	17	1.0
									Oceania								
Australia	6	1.4				Australia	3	1.0	Australia	2	0.2	Australia	1	0.9	Australia	3	0.2
									New Guinea	1	0.1				New Zealand	1	0.1

Table 9

Distribution of Traffic Source of All Master Visitors in each cluster

Cluster 1 Interested-HP			Cluster 2 Scholarship-Driven			Cluster 3 MA-P			Cluster 4 Attention			Cluster 5 Desired-HP			Cluster 6 Interested		
Source	N	%	Source	N	%	Source	N	%	Source	N	%	Source	N	%	Source	N	%
Google/organic	224	53.2	Google/organic	82	59.9	Google/Organic	11	3.6	Google/organic	253	29.9	Google/organic	91	81.3	Google/organic	200	11.2
Quick link	82	19.5	Google/cpc	46	33.6	Google/cpc	1	0.3	Google/cpc	96	11.3	E-mail	22	19.6	Google/cpc	70	3.9
Google/cpc	31	7.4	Quick link	13	9.5	Direct	1	0.3	Quick link	38	4.5	Quick link	21	18.8	E-mail	50	2.8
Direct	19	4.5	Direct	6	4.4	E-mail	1	0.3	Bing	27	3.2	Facebook	12	10.7	Direct	40	2.2
E-mail	18	4.3	Facebook	4	2.9	Baidu	0	0.0	Direct	20	2.4	Google/cpc	10	8.9	Yahoo	20	1.1
Facebook	10	2.4	Bing	3	2.2	Bing	0	0.0	Facebook	17	2.0	Direct	10	8.9	Baidu	10	0.6
Bing	9	2.1	Yahoo	3	2.2	Yahoo	0	0.0	E-mail	11	1.3	Bing	2	1.8	Bing	10	0.6
Yahoo	4	1.0	E-mail	3	2.2	Facebook	0	0.0	Yahoo	4	0.5	Yahoo	1	0.9	Facebook	10	0.6
Baidu	3	0.7	Baidu	0	0.0	Quick link	0	0.0	Baidu	3	0.4	Baidu	0	0.0	Quick link	0	0

Table 10

Distribution of preferred Device Type of All Master Visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Interested-HP		Scholarship-Driven		MA-P		Attention		Desired-HP		Interested	
Device Type	N	%	N	%	N	%	N	%	N	%	N	%
Desktop	258	61.3	84	61.3	11	3.6	313	37.0	94	83.9	200	11.2
Mobile	74	17.6	54	39.4	3	1.0	143	16.9	15	13.4	80	4.5
Tablet	12	2.9	4	2.9	0	0.0	14	1.7	1	0.9	0	0

4.3.2 Behavioural Profiling of Dutch Master Visitors

Dutch Master visitors are analysed in order to discover their behaviour and identify differences compared to All Master Visitors. As determined, an appropriate number of clusters for Dutch Master Visitors is six. Therefore, the K-modes algorithm distributes all Dutch Master visitors (N=412) into 6 clusters. In Table 11, Cluster 4 is the largest cluster and contains 36.41% of Dutch Master Visitors. The second largest is cluster 6 (33.25%), followed by cluster 2 (12.86%). Cluster 4, 6, and 2 contain about 82% of all Dutch Master Visitors. The remaining 18% is distributed in cluster 5 (7.52%), cluster 1 (5.83%), and cluster 3 (4.13%).

Table 11

Distribution of Dutch Master Visitors in each cluster

	N	%
Cluster 1	24	5.83
Cluster 2	53	12.86
Cluster 3	17	4.13
Cluster 4	150	36.41
Cluster 5	31	7.52
Cluster 6	137	33.25
Total	412	100

In the 1st cluster of Dutch Master Visitors in Table 12, 70.8% converted by submitting their application. Members of this cluster have the highest conversion percentage in comparison to all other clusters. The most Prominent behavioural attributes are Eligibility Check (95.8%), Educational Brochure Request (91.7%), and Managed CTA click (79.2%). Educational Brochure Requests implies that visitors are aware of UT studies and conducting the Eligibility Check demonstrates a visitors interest to study at the UT. Additionally, the high conversion rate implies that members of this cluster can be considered high potential prospects. Furthermore, a large proportion of Dutch visitors have Registered for an Open Day (29.2%), asked a question via Web form (25.0%), and Requested Student for a Day (20.8%). Additionally, 16.7% downloaded some form of PDF and 16.7% considered the FAQ page. This distinctive pattern of behaviour demonstrates that members in cluster 1 are highly interested in UT offerings. Approximately 8 out of 10 visitors manifested 4 of 10 behaviours and about 3 out of 10 visitors manifested 7 of the 10 behaviours. Hence, it could explain the high conversion percentage. These characteristics resemble the Desire stage of the AIDA-model. As cited in Wijaya (2012) the desire stage is characterized by the development of a favourable attitude towards the institute (i.e., affection). Therefore, cluster 1 can be labelled as *Desired High Potential Prospects* or *Desired-HP*. Popular studies are Health Sciences (20.8%), closely followed by Business Administration (16.7%), and Applied Mathematics (12.5%) (Table 13). Lastly, 95.8% of the traffic comes through Google/organic (Table 14) and the preferred device type is a Desktop (91.7%) followed by mobile (25%) (Table 15).

In the 2nd cluster, 1.9% converted on Osiris Application Submitted. A distinctive characteristic are the Questions via Web form which is manifested by 100% of the cluster members. Furthermore, the Eligibility Check is taken by 37.7%, 15.1% requested student for a day, and 11.3% downloaded an Educational Brochure. All members have asked a question via Web form which can imply that the visitors could not find all the information they were looking for on the website. Additionally, the fair amount of visitors that registered for an open day may compensate for the missing information on the website and visitors might prefer to visit an open day over downloading educational brochures. However, the most prominent are Questions via Web form and Eligibility Check. It resembles visitors who want to acquire more detailed information and are interested whether they meet the minimum requirements for UT studies. These behaviours are related to the interest stage of the AIDA-model. However, PDF downloads or Educational Brochure requests are less prominent which decreases their awareness. Therefore, this cluster can be labelled as *Moderately Aware Potential Prospects* or *MA-P*. Popular studies are Business Administration (15.1%), Health Sciences (11.3%), and Electrical Engineering (12.5%). The majority of traffic comes through Google/organic (20.8) and the preferred device type is a desktop (19.8%).

In the 3rd cluster, 29.4% converted on Osiris Application Submitted. This clusters includes the second highest conversion score. Prominent attributes are the Eligibility Check (100%), Educational Brochure Request (88.2%), Questions via Web form (76.5%), and PDF download (64.7%). This cluster manifests the highest percentage of PDF downloads. A PDF download entails that a visitor downloaded some kind of PDF that is non-study related. This can be in terms of financial information, a ground plan, or a UT catalogue. The high percentages of PDF downloads and Educational Brochure requests demonstrates that visitors want to obtain more information which in turn raises their awareness and interest. Additionally, all visitors considered the Eligibility Check which confirms their high level of interest. Furthermore, a substantial proportion asked a question via Web form which implies they are curious to acquire additional information about the UT that was not available on the website. About 17.6% conducted an Open Day Registration, Requested Student for a Day, and considered the FAQ page. However, these are less prominent compared to the highest converting visitors in cluster 1, resulting in a lower conversion rate. Hence, cluster 3 can be labelled as *Interested High*

Potential Prospects or *Interested-HP*. Popular are Communication Studies (29.4%), Civil Engineering and Management (17.6%), and Construction Management and Engineering (17.6%). Most traffic comes via Google/organic (70.6%) and Direct visits (11.8%). Lastly, the preferred device type is a Desktop (64.7%).

In the 4th cluster, 5.3% converted by submitting their application. It includes the third highest conversion percentage in comparison to all clusters. Cluster 4 is the largest cluster in terms of size (36.41%). The most prominent behaviour is the Eligibility Check that has been conducted by all cluster members (100%). Additionally, 12.7% requested an Educational Brochure, 5.3% requested Student for a Day, and 4.0% clicked on a Managed CTA. The presence of the remaining attributes are negligible. As the majority only performed one or two behaviours it can potentially explain why this cluster includes a lower conversion percentage compared to cluster 1 and 3. A small proportion in cluster 4 requested an Educational Brochure. This indicates that members of cluster 4 could be moderately aware about the UT study programmes. However, all members conducted the Eligibility Check which implies that members of this cluster are interested to see whether they meet the minimum requirements for the study programme they are interested in. Hence, it can be assumed that visitors are interested in UT studies. There is small proportion of visitors that conduct any kind of other behaviour to acquire information about the UT. Hence, it is highly probable that visitors in this cluster obtained information about the UT from sources other than the website. As mentioned, potential sources are a positive spread of Word of Mouth (WOM), Electronic Word of Mouth (eWOM) through social media channels, and online and offline advertising. It can be argued that members of cluster 4 became aware by external sources, some by Educational Brochures, and are interested if they are eligible for UT studies. These characteristics resemble the interest stage of the AIDA-model which is associated with customers who have developed an affiliation for the institute to some degree (Wijaya, 2012). Hence, cluster 4 is labelled as the *interested* group. Popular studies are Communication Studies (11.3%), Psychology (8.7%), and Sustainable Energy Technology (8.0%).Furthermore, 12.4% visits via Google/organic and the preferred device type is a Desktop (15.3%).

In the 5th cluster, 0% of visitors converted on Osiris Application Submitted. Prominent attributes are Educational Brochure Request (100%), Managed CTA click (90.3%), and Open Day Registration (54.8%). Additionally, 16.1% requested to be student for a day. The presence of the remaining behaviours are less prominent to make any kind of distinction. These characteristics resemble visitors with an interest to acquire more information as well as experiencing the student life at the UT. The distinctive characteristic of this group is that all visitors requested an Educational Brochure which raises their awareness. Additionally, the majority of this group visits the website via Google/organic (81.1%) and nearly all visitors have clicked on a Managed Call-to-Action. This implies that visitors came across UT advertisements and downloaded various forms of Educational Brochures for more detailed information. As the remaining behaviours are rarely manifested it could potentially explain the low conversion of Osiris Application Submitted. The latter is supported by the possibility that they require more information before making their decision by visiting Open Days or as student for a day. Therefore, it can be hypothesized that members of cluster 5 are interested in UT studies but in addition to the brochures they want to experience the UT. Members of cluster 5 manifest experience oriented behaviours which is distinctive compared to other clusters. Hence, cluster 5 is labelled as *Experience-Driven*. Popular studies are Health Sciences (51.6%), Biomedical Engineering (6.5%), and Computer Science (6.5%). The majority of the traffic comes via Google/organic (81.1%) and the preferred device type is a Desktop (80.6%).

In the 6th cluster of Table 12, 4.4% converted on Osiris Application Submitted. This cluster is characterized by Educational Brochure Requests conducted by all visitors (100%) which raises their awareness about UT studies. Furthermore, 7.3% downloaded some form of PDF.

The presence of the remaining behavioural attributes are negligible. The majority of this group visits the website via Google/organic (51.8%). It could be hypothesized that the visitors became aware by a positive spread of Word of Mouth (WOM), eWOM, or advertisements and used Google to obtain more information. Therefore, they have downloaded various forms of Educational Brochures. These characteristics resemble visitors that are in the Attention stage of the AIDA-model. In this stage customers become aware of the products and services and seek to acquire more information. Hence, this stage is associated with cognition and rational knowledge seeking (Wijaya, 2012; Rawal, 2013). As a result, this cluster is labelled as the *Attention* group. Popular studies are Health Sciences (11.7%), Business Administration (9.5%), and Environmental and Energy Management (9.5%). Lastly, the majority prefers Google/organic (51.8%) and the preferred device type is a Desktop (48.2%).

Table 12

Distribution of Behavioural Attributes of Dutch Master Visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Desired-HP		MA-P		Interested-HP		Interested		Experience-Driven		Attention	
Behavioural Attributes	N	%	N	%	N	%	N	%	N	%	N	%
Educational Brochure Request	22	91.7	6	11.3	15	88.2	19	12.7	31	100.0	137	100.0
Eligibility Check	23	95.8	20	37.7	17	100.0	150	100.0	2	6.5	0	0.0
Questions via Web form	6	25.0	53	100.0	13	76.5	0	0.0	1	3.2	0	0.0
Request Student for a Day	5	20.8	1	1.9	3	17.6	3	2.0	5	16.1	5	3.6
Osiris Application Submitted	17	70.8	1	1.9	5	29.4	8	5.3	0	0.0	6	4.4
Managed CTA Click	19	79.2	5	9.4	2	11.8	6	4.0	28	90.3	0	0.0
PDF Download	4	16.7	2	3.8	11	64.7	3	2.0	2	6.5	10	7.3
Scholarship Finder	1	4.2	1	1.9	3	17.6	0	0.0	1	3.2	4	2.9
Open Day Registration	7	29.2	8	15.1	3	17.6	8	5.3	17	54.8	0	0.0
FAQs	4	16.7	1	1.9	3	17.6	0	0.0	2	6.5	3	2.2

Table 13

Top Three Study Programmes of Dutch Master Visitors in each cluster

Cluster 1 Desired-HP			Cluster 2 MA-P			Cluster 3 Interested-HP			Cluster 4 Interested			Cluster 5 Experience-Driven			Cluster 6 Attention		
Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%
HS	5	20.8	BA	8	15.1	CS	5	29.4	CS	17	11.3	HS	16	51.6	HS	16	11.7
BA	4	16.7	HS	6	11.3	CEM	3	17.6	P	13	8.7	BE	2	6.5	BA	13	9.5
AM	3	12.5	EE	5	9.4	CME	3	17.6	SET	12	8.0	COMPS	2	6.5	EEM	13	9.5

Note. Explanations of abbreviations are available on p.3. Complete list of studies per cluster are available in Appendix I.

Table 14

Distribution of Traffic Source of Dutch Master Visitors in each cluster

Cluster 1 Desired-HP			Cluster 2 MA-P			Cluster 3 Interested-HP			Cluster 4 Interested			Cluster 5 Experience-Driven			Cluster 6 Attention		
Source	N	%	Source	N	%	Source	N	%	Source	N	%	Source	N	%	Source	N	%
Google/organic	23	95.8	Google/organic	11	20.8	Google/organic	12	70.6	Google/organic	19	12.7	Google/organic	27	87.1	Google/organic	71	51.8
Facebook	6	25.0	Google/cpc	2	3.8	Direct	2	11.8	Google/cpc	7	4.7	Google/cpc	10	32.3	Quick link	13	9.5
E-mail	6	25.0	Direct	1	1.9	Quick link	2	11.8	Direct	4	2.7	Quick link	9	29.0	Google/cpc	7	5.1
Quick link	6	25.0	Bing	1	1.9	Google/cpc	1	5.9	E-mail	3	2.0	E-mail	2	6.5	Bing	5	3.6
Google/cpc	4	16.7	E-mail	1	1.9	E-mail	1	5.9	Quick link	3	2.0	Direct	1	3.2	E-mail	5	3.6
Direct	3	12.5							Facebook	2	1.3	Facebook	1	3.2	Facebook	2	1.5
Bing	1	4.2							Bing	1	0.7						

Table 15
Distribution of preferred Device Type of Dutch Master Visitors in each cluster

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Desired-HP		MA-P		Interested-HP		Interested		Experience-Driven		Attention	
Device Type	N	%	N	%	N	%	N	%	N	%	N	%
Desktop	22	91.7	10	19.8	11	64.7	23	15.3	25	80.6	66	48.2
Mobile	6	25.0	5	9.4	3	17.6	6	4.0	6	19.4	18	13.1
Tablet	1	4.2					1	0.7	1	3.2	3	2.2

4.4 Comparison Analysis

Previous sections indicated that six behavioural profiles exist among website visitors interested in Master studies at the University of Twente. The discovered profiles are labelled as Interested High Potential Prospects, Scholarship-Driven Prospects, Moderately Aware Potential Prospects, Attention Prospects, Desired High Potential Prospects, and Interested Prospects for All Master Visitors. The behavioural profiles of Dutch Master visitors are labelled as Desired High Potential Prospects, Moderately Aware Potential Prospects, Interested High Potential Prospects, Experience-Driven Prospects, and Attention Prospects. Each of the profiles is comprised of a distinctive behavioural pattern which is described in terms of its most prominent behavioural attributes. However, a more detailed picture is required in order to identify similarities and dissimilarities in behavioural attributes between All Master Visitors and Dutch Master Visitors. Furthermore, focusing on prominent behavioural attributes may not provide a true picture of the profiles and its underlying structures. Therefore, this section aims to evaluate the findings of the previous sections and uncover details that may only become visible by means of a comparison analysis. First, a comparison is given including the distribution of visitors per profile. Secondly, a comparison of the distribution of behavioural attributes per profile is depicted. Lastly, the distribution of Traffic Source, Study Programmes, and Device Type is compared for each profile.

4.4.1 Comparison of Distribution of Visitors per Profile

Figure 8 depicts the distribution of All Master visitors and Dutch Master visitors for each of the six behavioural profiles. At a macro-level it can be observed that for both groups the Interested profile is the largest followed by the Attention profile. However, the profiles are significantly different in terms of proportions. The proportion of All Master visitors in the Interested profile (50%) is larger compared to the Interested Profile of Dutch Master visitors (36%). Furthermore, the Desired-HP profile of All Master Visitors (3%) is smaller compared to the proportion of the Desired-HP profile of Dutch Master Visitors (6%). The interested-HP profile of All Master Visitors (12%) is substantially larger than the Interested-HP profile of Dutch Master Visitors (4%). In contrast, the MA-P profile of All Master visitors (8%) is substantially smaller than the MA-P profile of Dutch Master visitors (13%). Hence, the proportion of Desired High Potential prospects is larger for Dutch Master Visitors and the proportion of Interested high potential prospects is larger for All Master visitors. Finally, both groups have a profile including Scholarship-Driven for All Master visitors (4%) and Experience-Driven for Dutch Master visitors (8%). In brief, the largest profiles are the Interested and Attention profiles for both groups. All profiles are distinct in terms of proportions and both groups have one unique behavioural profile.

Figure 8. Distribution of Visitors per Profile

All Master Visitors	%	Dutch Master Visitors	%
Interested	49.5	Interested	36.41
Attention	23.4	Attention	33.25
Interested-HP	11.7	MA-P	12.86
MA-P	8.5	Experience-Driven	7.52
Scholarship-Driven	3.8	Desired-HP	5.83
Desired-HP	3.1	Interested-HP	4.13
Total	100.0	Total	100

4.4.2 Comparison of Behavioural Attributes per Profile

Figure 9 depicts the distribution of behavioural attributes of All Master Visitors and Figure 10 of Dutch Master Visitors per profile. Each profile is marked at 50% with a dashed line in order to make interpretations less complicated. A behavioural attribute manifested more than 50% is considered to be a *dominant* or *macro-behaviour* and attributes that are manifested less than 50% are considered to be *non-dominant* or *micro-behaviours*. The *Attention* and *Interested* profiles are the largest profiles in terms of size and similar for both groups at a macro-level. In both groups, the Educational Brochure Requests are a dominant behaviour (100%) for the *Attention* profile, and the Eligibility Check is dominant (100%) in the *Interested* profile. Hence, it can be hypothesized that at a macro-level the Attention and Interested profiles are similar across countries. Specifically, this is assumed to be valid between All Master visitors and Dutch Master visitors.

The *MA-P profiles* are equal in terms of *questions via web form* (100%) which can indicate certain information was not found on the website. However, profiles differ in the proportion of All Master visitors who conducted the Eligibility Check. It can be hypothesized that All Master visitors in the MA-P profile are more likely to conduct the Eligibility Check (67.3%) to see whether they meet the minimum requirements for UT studies compared to Dutch Master visitors (37.7%).

The *Interested-HP profiles* are different in terms of the proportions in macro and micro-behaviours between both groups. Among both profiles Educational Brochure Request and Eligibility Check are dominant. An important difference is that the Interested-HP profile of Dutch Master visitors manifest Questions via Web form (76.5%) and PDF downloads (64.7%) as additional dominant attributes. These attributes can contribute to their awareness (i.e., knowledge) and Interest (i.e., affection). Hence, it could explain the higher conversion rate on Osiris Application Submitted by Dutch Master visitors (29.4%) compared to All Master visitors (14%). Furthermore, the micro-behaviour Managed CTA click appears to be more manifested by All Master visitors (26.1%) compared to Dutch Master visitors (11.8%). For the Interested-HP profile, it could be hypothesized that *Questions via Web form* and *PDF downloads* have a positive influence on Osiris Application submitted for Dutch Master visitors in this profile.

The *Desired-HP profiles* include the highest converting UT website visitors. All Master visitors manifested Osiris Application Submitted by 78.6% and Dutch Master visitors by 70.8%. Both groups manifest the *Eligibility check, Educational Brochure, and Managed CTA click* as dominant attributes. However, PDF download is an additional dominant attribute for the Desired-HP profile of All Master Visitors whereas for Dutch Master visitors is it non-dominant. Hence, it could explain the difference of approximately 8% on Osiris Application Submitted conversions. Similar to the Interest-HP profiles, it could be hypothesized that PDF-download has a positive influence on Osiris Applications Submitted. Furthermore, all Master

visitors make more use of the *Scholarship finder* compared to Dutch Master visitors in the Desired-HP profiles. Hence, it can be argued that Scholarship availability is an important factor when considering to study at the UT for All Master Visitors. Moreover, *registration for Open Days* and *Requesting Student for a Day* is more prominent among Dutch Master visitors in the Desired-HP profile. Therefore, it can be hypothesized that Dutch Master visitors are more likely to consider Open Days or Student for a Day before converting and All Master visitors are more likely to consider website materials such as PDF-downloads and Scholarship Finder.

The *Scholarship-Driven* uniquely characterizes All Master visitors compared to Dutch Master visitors. It can be argued that *Scholarship availability* is an important factor when considering to study at the UT for All Master visitors. Hence, it can be hypothesized that their conversion rate is possibly influenced by Scholarship Availability.

The *Experience-Driven* profile uniquely characterizes Dutch Master visitors in terms of *Open Day registrations* and *Requesting Student for a Day*. In addition to Educational Brochure Requests Dutch Master visitors want to acquire information by experiencing the UT by visiting Open Days or being a Student for a Day. Therefore, it can be hypothesized that the conversion rate of Dutch Master visitors is highly influenced by Open Day Registrations and Requesting to be Student for a Day. Lastly, the Scholarship-Driven and Experience-Driven profiles all manifested a Managed CTA click (100%) which implies that visitors came across UT advertisements and downloaded various forms of Educational Brochures (100%) to acquire more detailed information.

Figure 9. Distribution of Behavioural Attributes of All Master Visitors

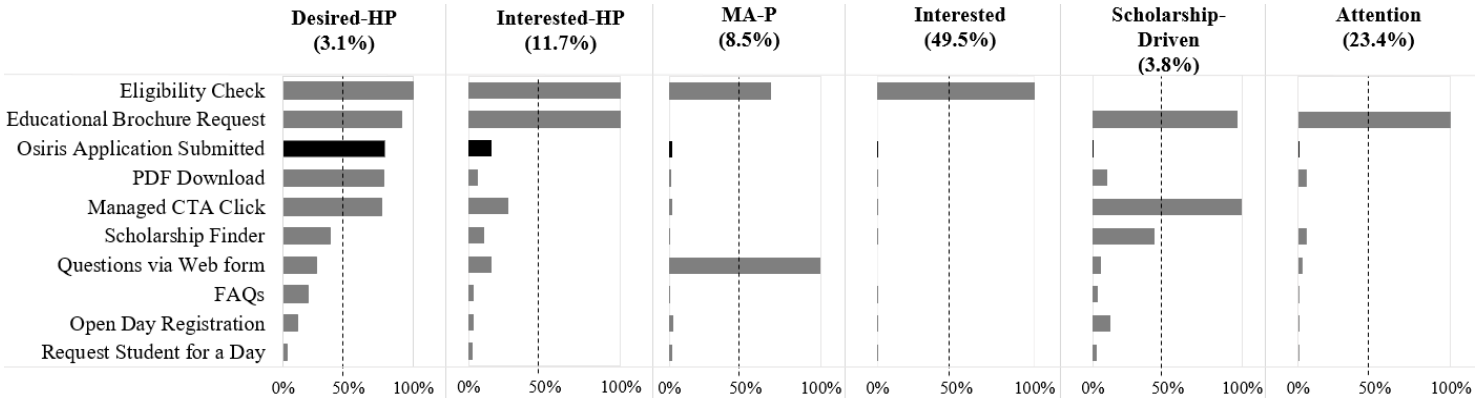
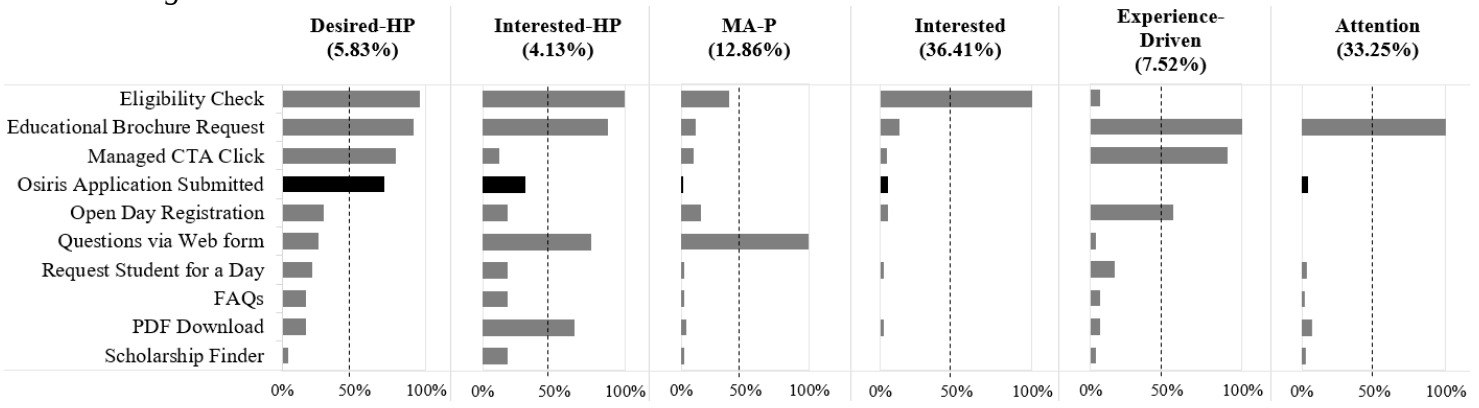


Figure 10. Distribution of Behavioural Attributes of Dutch Master Visitors



4.4.3 Comparison of Traffic Source per Profile

Figure 11 depicts the distribution of Traffic Sources for All Master visitors and Figure 12 for Dutch Master visitors. It can be concluded the majority of visitors enter the UT website via Google/organic, followed by Google/cpc, Quick links, and Direct visits. Interestingly, the Desired-HP profile is the only profile that is distinct in terms of E-mail and Facebook referrals which are significantly higher compared to other profiles. Hence, it could explain why Managed CTA click is one of the dominant behavioural attributes in the Desired-HP profiles. It can be argued that high converting visitors prefer Quick link, E-mail, and Facebook after Google/organic whereas lower converting visitors prefer Google/organic, Quick link, and direct visits. It can be concluded that dominant Traffic Sources are in similar order between profiles (i.e., macro-level) but non-dominant Traffic Sources differ in terms of proportions (i.e., metric-level) between All Master visitors and Dutch Master visitors.

Figure 11. Distribution of Traffic Source of All Master Visitors

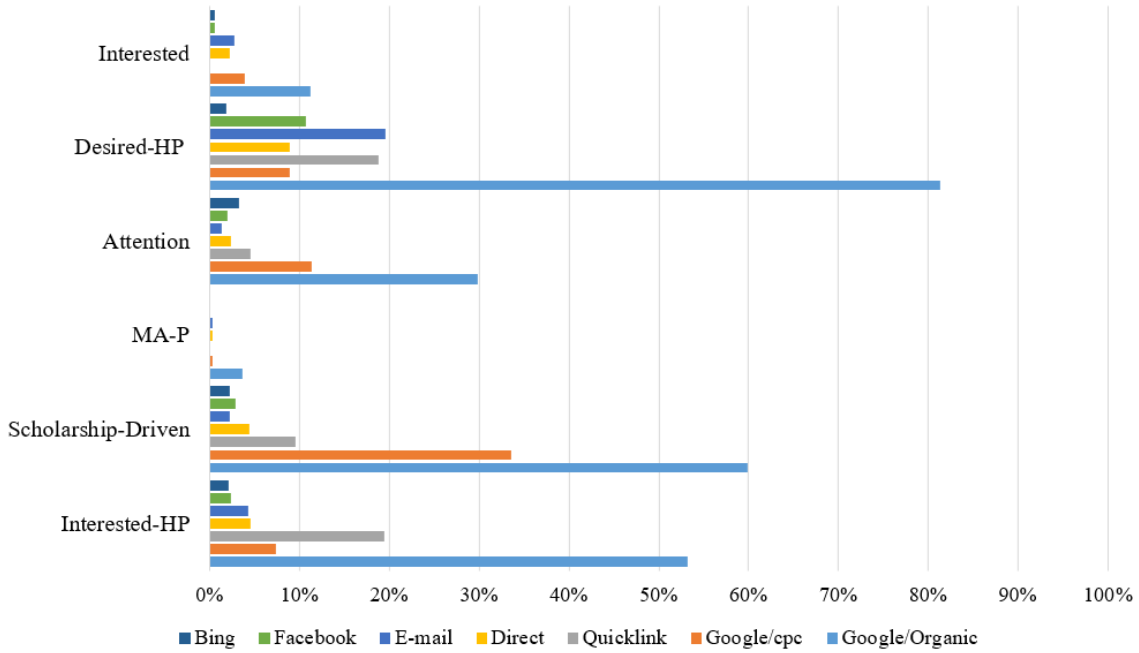
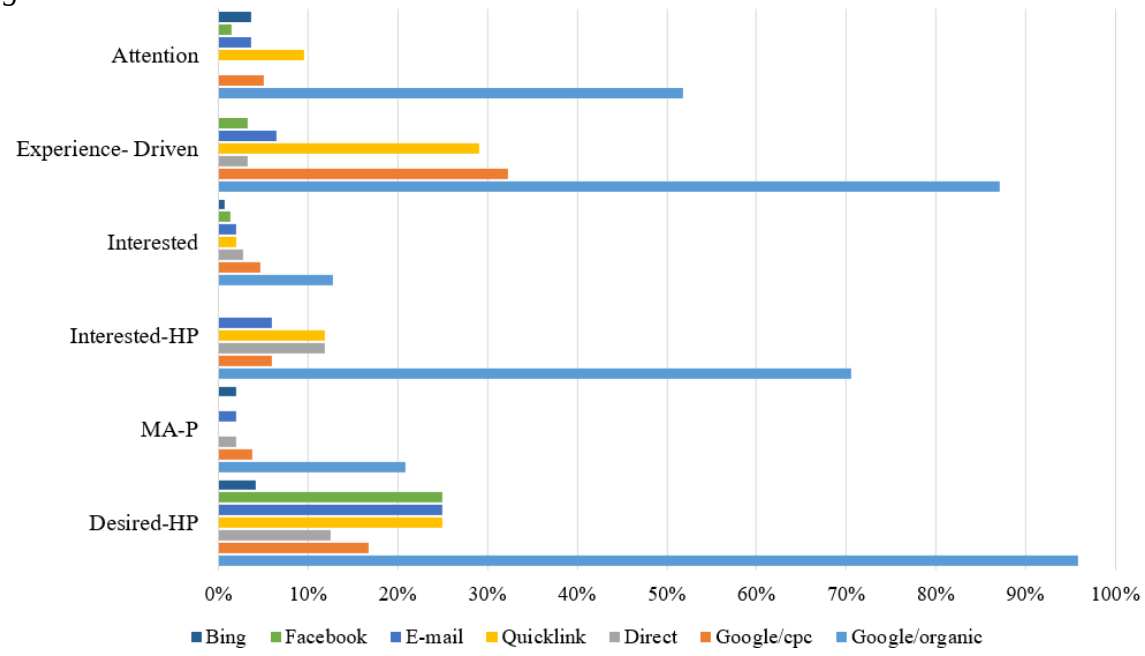


Figure 12. Distribution of Traffic Source of Dutch Master Visitors



4.4.4 Comparison of Preferred Device Type per Profile

The preferred device type of All Master visitors is depicted in Figure 13 and for Dutch Master visitors in Figure 14. In general, the majority of visitors prefer a Desktop followed by a Mobile and Tablet. This behaviour is similar between all profiles and across both groups. In both groups the Desired-HP profile includes the largest proportion of visitors who prefer a Desktop compared to all other profiles in both groups. However, it appears that All Master visitors make more use of Mobile devices compared to Dutch Master visitors. Furthermore, profiles of high converting visitors are more likely to use a Desktop compared to profiles of low converting visitors. It can be argued that a Desktop could be more convenient for visitors who manifest multiple dominant behaviours (e.g., high potential prospects). It can be hypothesized that the order of preferred device types are similar in all profiles and across groups but at a micro-level (metric) it varies between All Master visitors and Dutch Master visitors.

Figure 13. Distribution of Preferred Device Type of All Master Visitors

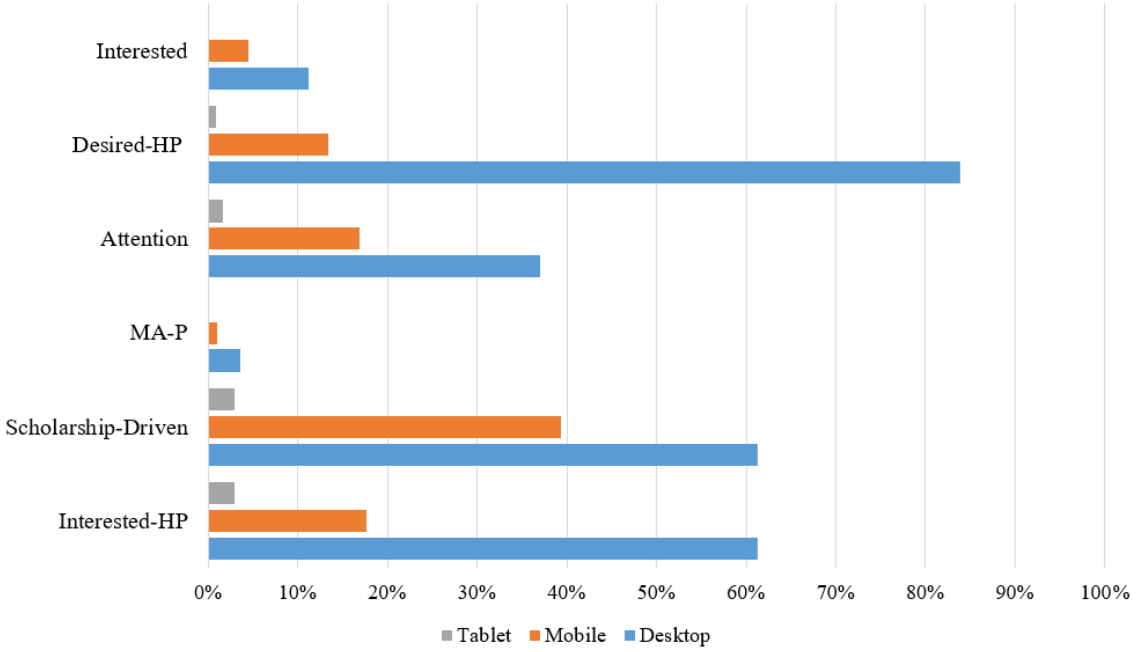
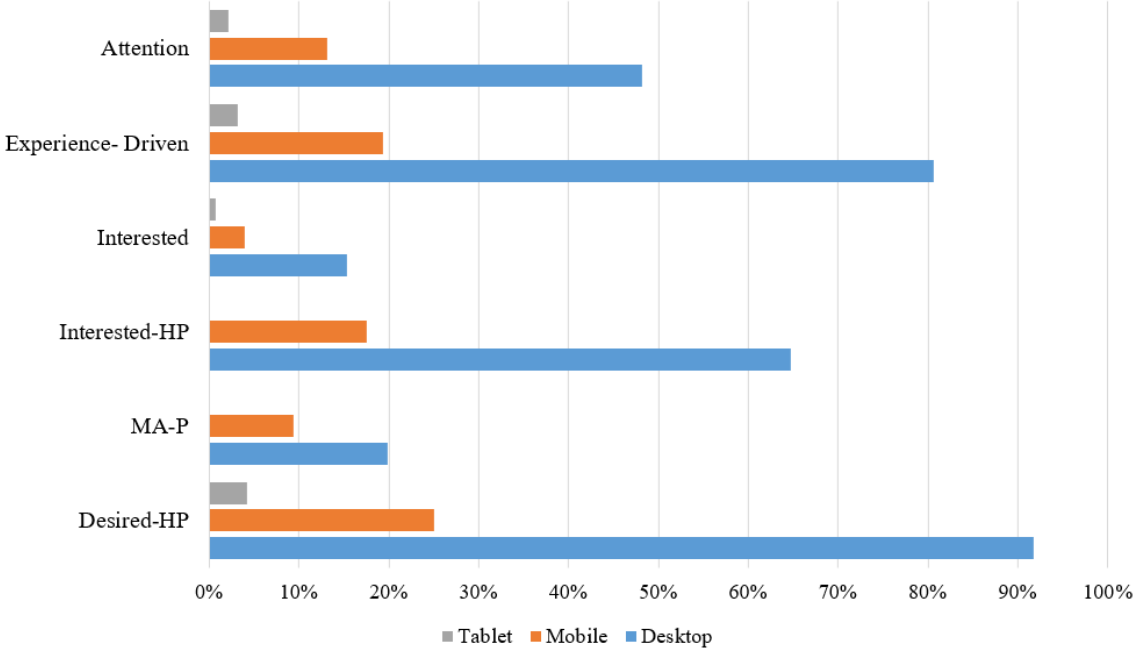


Figure 14. Distribution of Preferred Device Type of Dutch Master Visitors



4.4.5 Comparison of Study Programmes per Profile

The top three study programmes per profile of All Master visitors are depicted in Table 16 and for Dutch Master visitors in Table 17. The most recurrent studies among all profiles of All Master Visitors are Mechanical Engineering (ME), Civil Engineering and Management (CEM), and Sustainable Energy Technology (SET), followed by Electrical Engineering (EE), Spatial Engineering (SE), and Computer Science (CPS). The most recurrent studies among Dutch Master visitors are Business Administration (BA), Health Sciences (HS), and Communication Studies (CS). The results indicate that Dutch Master visitors are more likely to be interested in studies of Behavioural, Management, and Social Sciences (BMS) whereas All Master visitors are more likely to be interested in studies of Engineering Technology (ET). For example, All Master visitors of the Desired-HP profile are more interested in CEM (18.8%), SET (17.0%), and ME (15.2%). In contrast, Dutch Master visitors are interested in HS (20.8%), BA (16.7%), and AM (12.5%). It can be hypothesized that the interest in UT Master studies differ across countries on a macro-level (ordinal) and micro-level (metric). At least, this hypothesis could be valid between All Master visitors and Dutch Master visitors.

Table 16

Popular Study Programmes of All Master Visitors in each Profile

Cluster 1 Interested-HP			Cluster 2 Scholarship-Driven			Cluster 3 MA-P			Cluster 4 Attention			Cluster 5 Desired-HP			Cluster 6 Interested		
Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%
SET	67	15.9	HS	17	12.4	EE	30	9.8	SE	108	12.8	CEM	21	18.8	ME	195	10.9
ME	48	11.4	SE	14	10.2	BA	28	9.2	GISEO	84	9.9	SET	19	17.0	CPS	170	9.5
CEM	47	11.2	CEM	13	9.5	CPS	27	8.8	EEM	83	9.8	ME	17	15.2	EE	170	9.5

Table 17

Popular Study Programmes of Dutch Master Visitors in each cluster

Cluster 1 Desired-HP			Cluster 2 MA-P			Cluster 3 Interested-HP			Cluster 4 Interested			Cluster 5 Experience-Driven			Cluster 6 Attention		
Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%
HS	5	20.8	BA	8	15.1	CS	5	29.4	CS	17	11.3	HS	16	51.6	HS	16	11.7
BA	4	16.7	HS	6	11.3	CEM	3	17.6	P	13	8.7	BE	2	6.5	BA	13	9.5
AM	3	12.5	EE	5	9.4	CME	3	17.6	SET	12	8.0	CPS	2	6.5	EEM	13	9.5

4.5 Clustering Validation

Assessing the cluster validity is an important aspect in clustering analysis. As a result, conclusions based on the discovered profiles become more robust. Following chapter 3, the Silhouette score and Cross-Validation are performed.

4.5.1 Silhouette Score

As described in chapter 3, the silhouette's test evaluates the cluster homogeneity (i.e., unity within-clusters) by measuring how closely each cluster member is located to its profile centroid. Furthermore, it measures the average distance between clusters and the degree to which the observations are well structured. A score of 1 implies that the cohesion within the clusters is quite good. A score closer to -1 indicates that the cohesion within the clusters is poor and not as valid. The results are presented in Table 18.

The silhouette score of All Master visitors is 0.76 which indicates the clusters are cohesive. The score for Dutch Master students is 0.56. Although the score is not as high compared to All Master Visitors it still indicative of a reliable clustering solution as the scores are closer to 1 than 0 or -1.

Table 18
Silhouette Score of All Master Visitors and Dutch Master Visitors

All Master Visitors	Dutch Master Visitors
0.763427236482932	0.562356414382147

4.5.2 Cross-Validation

Cross-validation is performed for All Master Visitors and Dutch Master Visitors by randomly splitting the original data into two sub-samples. The sub-samples consist of a training dataset (75%) and a test dataset (25%). The hold-out method is applied and used to evaluate the appropriateness of the chosen number of clusters as described in chapter 3. This is done by recalculating the number of clusters for the training dataset and the test dataset. Hence, the Hamming distance is recalculated for both sub-samples and the number of clusters are determined by hierarchical clustering. In addition, the relative loss of inertia is calculated to improve the interpretability. The size of the training dataset is 2709 and for the test dataset 903 for All Master Visitors. The size for Dutch Master Visitors is 309 for the training dataset and 103 for the test dataset. It is important to note that the hold-out method is a simple variation of cross-validation. It involves only a single run whereas more exhaustive methods run several times on multiple *k*-partitions. The *k* results are then averaged to obtain a single estimation. Hence, the hold-out method may include some variation depending on how the data is randomly split.

In Figure 14 the results of the training dataset are depicted for All Master Visitors. According to the relative loss of inertia method the best partition to cut the tree is 3 clusters (black) and the second-best is 5 clusters (grey). However, from the dendrogram it can be observed that 6 clusters are appropriate. The relative loss of inertia between 5 clusters and 6 clusters is approximately 0.001 which is negligible. Hence, 6 is an appropriate number of clusters according for the training dataset for All Master Visitors. In Figure 15, the results of the test dataset are depicted for All Master Visitors. The graph including the relative loss of inertia suggests that 4 clusters are the best partition to cut the tree and the second-best is 3 clusters. The dendrogam of the test dataset suggests that 6 is appropriate. The relative loss of inertia between 4 clusters and 6 clusters is approximately 0.003 which is negligible. Hence, it can be concluded that six is an appropriate number of clusters for All Master Visitors according to both the test dataset and training dataset.

The dendrogram and relative loss of inertia for the training dataset of Dutch Master Visitors are depicted in Figure 16 and the results of the test dataset can be found in Figure 17. It can be observed from the dendrogram and relative loss of inertia method that six is an appropriate number of clusters for both the training dataset and test dataset of Dutch Master Visitors.

Lastly, the Silhouette scores are recalculated for the training dataset and test dataset for both groups. For All Master Visitors the Silhouette score is 0.51 for the training dataset and 0.60 for the test dataset. The Silhouette scores for Dutch Master Visitors is 0.50 for the training dataset and 0.48 for the test dataset. The lower score on the test dataset is acceptable considering that all previous analysis and calculations depict positive results and the score may vary depending on how the dataset was randomly split. Furthermore, the scores are similar in comparison to section 4.5.1 of the original data. Hence, it can be concluded that the discovered behavioural profiles are valid and robust.

Figure 14. Dendrogram and relative inertia loss All Master Visitors Training Dataset

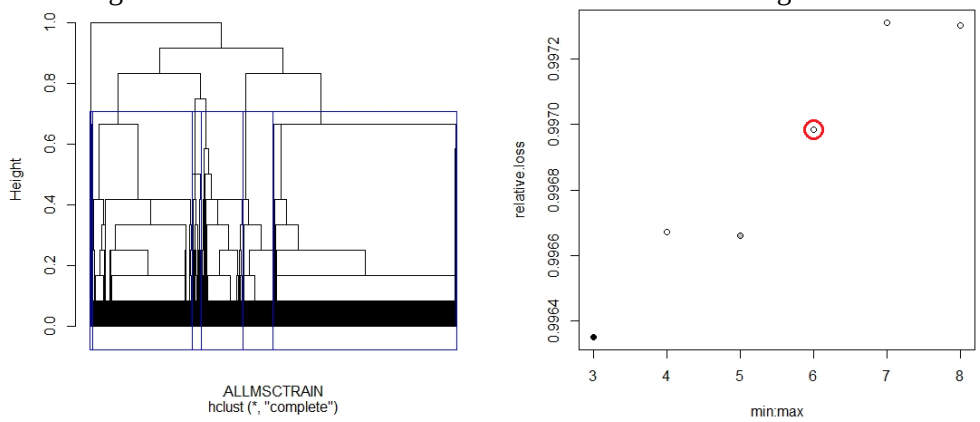


Figure 15. Dendrogram and relative inertia loss All Master Visitors Test Dataset

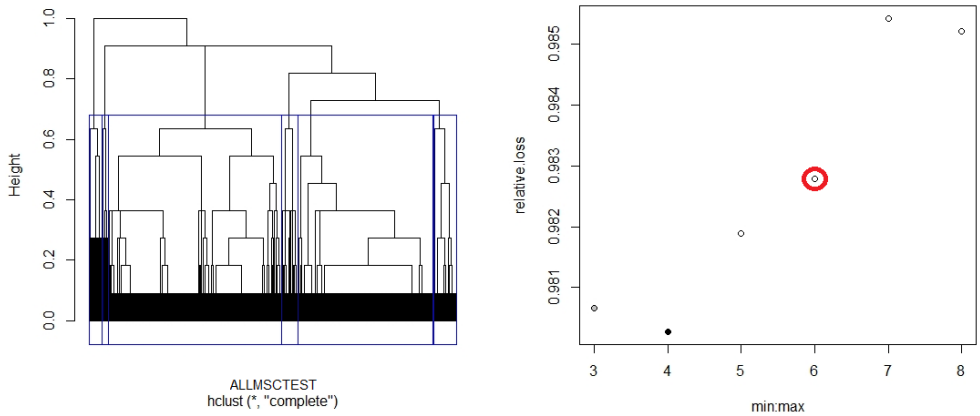


Figure 16. Dendrogram and relative inertia loss Dutch Master Visitors Training Dataset

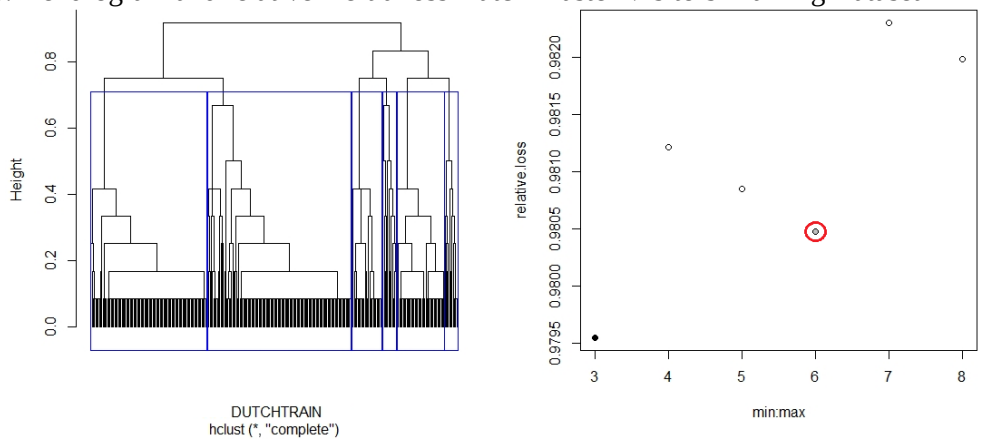
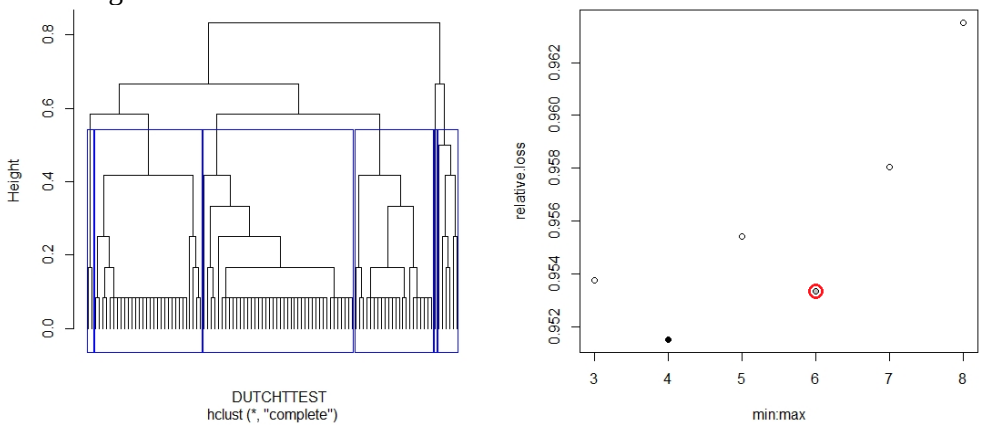


Figure 17. Dendrogram and relative inertia loss Dutch Master Visitors Test Dataset



5. DISCUSSION

A key competitive advantage for today's organizations is the availability of large amounts of data for the purpose of segmenting a customer base, offering tailored services, and extracting meaningful information provided by various data sources. However, organizations often have difficulties to extract knowledge from data and selecting appropriate ML and User Profiling approaches. For instance, Dolnicar (2002) found that marketing departments lack a fundamental understanding on data-driven segmentation methodologies. Key issues were determining the number of clusters and which algorithm should be chosen. Moreover, the interpretability and understandability of data-driven segments is an important issue due to a lack of research and cases and increasingly complex segmentation bases (Boratto et al., 2016; Dolnicar, 2002). Furthermore, segmentation approaches failed to acknowledge how different types of user data and segmentation criteria may affect the quality of User Profiles. In addition, numerous approaches were available for numerical data but approaches for categorical or mixed data were not as prevalent or straightforward.

The *first objective* was to develop a methodology and a framework of UML algorithms for User Profiling with respect to their requirements regarding various data properties. The research question was: *What is an appropriate framework for outlining UML Algorithms for User Profiling?* Among others, literature was reviewed on ML algorithms, their requirements regarding data properties, and two-stage clustering.

As a result, a *framework* is proposed outlining various UML algorithms for User Profiling with respect to various data properties. It provides a two-stage clustering methodology for categorical, numerical, and mixed types of data with respect to the data size and data dimensionality. The first stage consists of an hierarchical or model-based procedure to determine the number of clusters. In the second stage, a non-hierarchical clustering procedure is applied for cluster refinement. The framework can support researchers and practitioners to determine which UML algorithms are appropriate for developing robust user profiles and segments for marketing purposes. Selecting a UML algorithm is highly dependent on the data type, data size, and data dimensionality as these properties have a significant effect on the quality and efficiency of the clustering procedure and solution and are therefore included in the framework (Fahad et al., 2014, Pandove et al., 2018, Dolnicar., 2002; Han, 2012; Larose, 2014). Prior research focused on the development, effectiveness (i.e., accuracy), and efficiency of various UML algorithms (e.g., Tamasauskas et al., 2012; Pandove et al., 2018, Huang, 1998; Park et al., 2009). However, none provided an outline as proposed in this paper. Moreover, the two-stage clustering approach alleviates the drawbacks of solely using a hierarchical or non-hierarchical algorithm resulting in more robust clustering solutions (e.g., Kuo et al., 2002; Mazanec & Strasser, 2000; Punj & Steward, 1983).

The *second objective* was to utilize the framework to discover high converting online behavioural profiles of Dutch website visitors interested in Master studies at the University of Twente (UT). The second research question was as follows: *What online behavioural profiles of Dutch website visitors interested in UT Master studies are most significant in terms of conversions?* As a result, *six behavioural profiles* were discovered among Dutch website visitors and All website visitors interested in UT Master studies. The profiles were labelled as Desired High Potential Prospects, Interested High Potential Prospects, Moderately Aware Prospects, Interested prospects, Experience-Driven prospects (Dutch Master Visitors), Scholarship-Driven prospects (All Master Visitors), and Attention prospects. All profiles are distinguishable by *macro-behaviours* and especially by *micro-behaviours*.

The *Desired-HP* profiles include the *highest converting* visitors in both groups. PDF-download is an additional macro-behaviour for the Desired-HP profile of All Master Visitors in contrast to Dutch Master Visitors. It could potentially explain the difference in conversions on Osiris Application Submitted by approximately 8%. Moreover, the Desired-HP profiles

indicate that All Master visitors make more use of the *Scholarship Finder* and *PDF-downloads* whereas Dutch Master visitors make more use of *Open Day Registrations* and *Request Student for a Day* in addition to the more commonly shared macro-behaviours. The *Interested-HP* profiles include the *second-highest converting* visitors. However, the Interested-HP profile of Dutch Master Visitors manifests Questions via Web Form and PDF-downloads as additional macro-behaviours in contrast to All Master visitors. As the conversion rates differ it could be argued that *Questions via Web Form* and *PDF downloads* have a positive influence on the conversion rate in the Interested-HP profile of Dutch Master visitors. Comparing high converting profiles with low converting profiles it appears that profiles with more than *three* macro-behaviours have higher conversion rates than profiles with less than three dominant behaviours. *Educational Brochure Request, Eligibility Check,* and *Questions via Web Form* are the most commonly manifested macro-behaviours among profiles. However, macro-behaviours can overlap between profiles whereas micro-behaviours consistently depict a unique pattern. Focusing on micro-behaviours can result in more effective and efficient marketing campaigns.

The distribution of *Traffic Source* and *Preferred Device Type* are similar between groups on a macro-level but at a micro-level it varies between between groups. The majority enters the UT website via Google/organic, followed by Google/cpc, Quick links, and Direct visits. However, higher converting visitors more often prefer *E-mail*, *Quick link* and *Facebook* as a traffic source whereas lower converting visitors prefer Google, Quick link, or direct visits. Moreover, Dutch Master visitors have a higher percentage in Facebook referrals and slightly more in Quick link and E-mail. *Study programmes* differ across groups (i.e., country) and profiles on both macro-level (ordinal) and micro-level (metric). It appears that Dutch Master visitors are more likely to be interested in *Behavioural, Management, and Social Sciences* (BMS) whereas All Master visitors appear to be more interested in studies of *Engineering Technology* (ET). However, confirmatory research is necessary to test this hypotheses.

Notably, the *experience-driven* profile uniquely characterizes *Dutch* Master visitors in terms of *Open Day registrations* and *Requesting Student for a Day*. Hence, the conversion rate of Dutch Master visitors is possibly influenced by Open Day Registrations and Requesting Student for a Day. The *Scholarship-Driven* profile uniquely characterizes All Master Visitors, making scholarship availability an important aspect to study at the UT. However, both profiles include Educational brochure request and Managed CTA as macro-behaviours. It appears that visitors in the scholarship and experience-driven profiles came across UT advertisements, downloaded various forms of educational brochures, and considered either Open Day Registrations or Scholarships. This potentially explains why Google/CPC and Quick link are prominent traffic sources right after Google/Organic in these profiles. The *Moderately Aware Profiles* are equal across both groups and characterized by visitors who all asked a *question via web form*. This can imply that visitors are interested but require more information which might not be available on the website. The UT M&C Department could apply text mining to study the kind of questions being asked to improve website content and marketing campaigns. The *Attention* and *Interested* Profiles are similar across groups. The interested profiles are the largest in terms of size but only manifested the *Eligibility Check*. It is assumed that instead of downloading educational brochures they raised their awareness or interest by a positive WOM, e-WOM, or advertisements before conducting the E-check.

Furthermore, a *model* is proposed which allows for a multi-criteria evaluation on different segmentation bases and User Profiling approaches that can guide researchers and practitioners to develop robust profiles and segments. It includes criteria which are essential for effective customer segmentation, different categories of customer attributes, and Implicit, Explicit, and Hybrid profiling approaches. Utilizing the hybrid approach can yield the most accurate, interpretable, and actionable segmentation results as it alleviates the drawbacks of solely using implicit or explicit user data (Dolnicar, 2008; Cufoglu, 2014). The majority of literature focused

on the development of User Profiling approaches and methodologies but did not consider the requirements for effective customer segmentation. Typically, they used one type of user data which was often explicit and numeric and limited to metrics as click through rate, time spent on page, or number of pages visited. For instance, Yan et al. (2009) segmented users based on their responses to advertisements. Results showed that click-through rates improved by 670 percent when using BT. Bhatnagar and Papatla (2001) segmented customers by using their search behaviour to present personalised ads. Targeting was based on keywords a consumer entered in a search engine. Another technique used was monitoring the clickstream on advertisements to measure and ad's effectiveness (Chen & Stallaert, 2014). Yao et al. (2010) used ML to identify purchasing and spending amounts to generate customer profiles. In contrast, this study used ML to discover profiles and understand what behaviours possibly contribute to a macro-conversion. A study of Rindfleish (2003) focused on profiling based on geo-demographic data of students and used it to measure the potential of market segments in the HE market. However, this study uses actual individual behavioural interactions of visitors with the website to discover user profiles (i.e., implicit data) in combination with explicit data.

5.1 Theoretical Implications

This paper proposed a *framework* outlining various UML Approaches for User Profiling with respect to important data properties. It provides two-stage clustering strategies whereby an appropriate combination can be selected for categorical, numerical and mixed types of data while considering the data size, and data dimensionality. In the first stage, the number of clusters is determined by hierarchical clustering. In the second stage, a non-hierarchical algorithm is applied for cluster refinement. Numerous approaches were found in academic literature for numerical data but approaches for categorical or mixed data were less prevalent and straightforward. Moreover, a two-stage approach can yield more robust results as it alleviates the drawbacks of solely using a hierarchical or non-hierarchical method (e.g., Kuo et al., 2002; Mazanec & Strasser, 2000; Punj & Steward, 1983). Furthermore, it was found that selecting an appropriate UML algorithm dependent on the data type, data size, and data dimensionality as these properties have a significant effect on the quality and efficiency of the clustering procedure and solution (Fahad et al., 2014, Pandove et al., 2018, Dolnicar., 2002; Han, 2012; Larose, 2014). Prior research focused on the development, effectiveness (i.e., accuracy), and efficiency of various UML algorithms (e.g., Tamasauskas et al., 2012; Pandove et al., 2018, Huang, 1998; Park et al., 2009). However, none provided an outline as proposed in this paper. The framework *contributes to literature* regarding approaches and methodologies for UML and data-driven segmentation in a marketing context.

A symmetric binary dataset was analysed in this study and some argued there is no best performing similarity measure (e.g., Boriah et al., 2014). However, after reviewing literature and conducting this research it was found that complete linkage with the hamming distance, followed by k-modes, can be performed on a symmetric binary dataset to obtain meaningful and robust results. Moreover, Dutt, Ismail, and Herawan (2017) conducted a systematic literature review from 1983 to 2016 on clustering algorithms and their applications in educational contexts. Results indicated that none of the studies used the combination of complete linkage and k-modes and considered only one type of user data. Hence, an important *contribution to literature*, especially in educational contexts, is the proposed combination of complete linkage followed by k-modes for User Profiling. Specifically, in the case of a small to moderate sized symmetric binary dataset with low dimensionality and considering both implicit and explicit user data.

Furthermore, a *model* is proposed which allows for a multi-criteria evaluation on different segmentation bases and User Profiling approaches which can guide researchers to develop robust profiles. It includes criteria which are essential for effective customer segmentation, different categories of customer attributes, and Implicit, Explicit, and Hybrid profiling

approaches. In contrast, the majority of literature focused on the development of User Profiling approaches and methodologies but did not consider the requirements for effective customer segmentation. Utilizing a hybrid approach can result in the most accurate, interpretable, and actionable segmentation results as it alleviates the drawbacks of solely using implicit or explicit user data (Dolnicar, 2008; Cufoglu, 2014).

Lastly, the six behavioural profiles were distinguishable by macro-behaviours and micro-behaviours. The results provide valuable insights for the UT M&C department to improve their marketing efforts and are indicative of a good performance by complete linkage and k-modes on a moderate sized and low dimensional symmetric binary dataset. Micro-behaviours allow to tailor marketing efforts and their discovery demonstrated that the proposed methods can generate profound insights.

5.2 Practical Implications

The framework and model can guide practitioners in selecting appropriate UML methods and combinations of customer attributes for User Profiling in a marketing context. Prior research found there is a lack of understanding among (university) marketing departments on fundamental data-driven segmentation methodologies. The framework can contribute to this problem by outlining various UML algorithms based on the characteristics of the dataset. Furthermore, the understandability – or interpretability – of data-driven segments is difficult due to increasingly complex segmentation bases and a lack of research and cases. The model can be considered to overcome this problem to some degree as it enables for a multi-criteria evaluation on different customer attributes and types of User Profiling based on criteria that are essential for effective segmentation. The hybrid approach can yield the most accurate, interpretable, and actionable segmentation results as it alleviates the drawbacks of solely using implicit or explicit user data.

Six behavioural profiles were discovered among Dutch website visitors and All website visitors interested in Master studies at the University of Twente. The profiles are distinguishable by *macro-behaviours* and *micro-behaviours*. However, macro-behaviours can overlap between profiles whereas micro-behaviours consistently depict a unique pattern. Developing marketing strategies based on more commonly shared macro-behaviours might negatively influence BT efforts whereas considering micro-behaviours can result in more tailored BT efforts. The *experience-driven* profile uniquely characterizes *Dutch* Master visitors and the *Scholarship-Driven* profile uniquely characterizes All Master Visitors. The *Desired-HP* profile followed by the *Interested-HP* profiles include the highest converting visitors. The Desired-HP profiles indicate that *All* Master visitors are more likely to be influenced by *Scholarship Finder* and *PDF-downloads* whereas *Dutch* Master visitors are more likely to be influenced by *Open Day Registrations* and *Request Student for a Day*. These (micro) behaviours can have a positive influence on the conversion rate of Osiris Application Submitted in addition to the more commonly shared and less distinctive macro-behaviours including *Educational Brochure request, Eligibility Check,* and *Questions via Web Form*.

Traffic Source and preferred *Device Type* are similar between groups on a macro-level but vary at a micro-level. *Study programmes* differ across groups and profiles on both macro-level (ordinal) and micro-level (metric). Dutch Master visitors appear to be more interested in studies of *Behavioural, Management, and Social Sciences* (BMS) whereas All Master visitors are more likely to be interested in studies of *Engineering Technology* (ET). However, confirmatory research is necessary to test this hypotheses. The majority visits the UT website via Google Organic, followed by Google/cpc, Quick links, and Direct visits. However, after Google the higher converting visitors prefer *E-mail*, *Quick link*, or *Facebook* whereas lower converting visitors prefer Google, Quick link, or direct visits. Major BT approaches to consider are Onsite BT and BT via Ad Networks. Most macro-behaviours are country independent and micro-behaviours are country dependent between Dutch Master visitors and All Master visitors.

Key managerial takeaways for the UT M&C department are: (1) Focus on micro-behaviours to tailor marketing campaigns, (2) Distinguish Experience-Driven and Scholarship-Driven prospects, (3) Target prospects similar, or to become similar, to the higher converting profiles Desired-HP and Interested-HP, (4) Further tailor marketing efforts by study programme, traffic source, and device type, and (5) Sequential analyses, Classification, and Text Mining can be utilized for future improvements. The latter is discussed in the next section.

5.3 Future Research and Research Limitations

This paper laid the foundation for future research on BT and (un)supervised Machine Learning. First, future research could conduct a *sequential analysis* of manifested behaviours. Finding out the sequence of manifested behaviours allows to motivate and nurture website visitors throughout their customer journey. Furthermore, it allows to identify behaviour of high potential prospects in an early stage of the conversion funnel. Next, future research could develop a *classification model*. Classification is a type of Supervised Machine Learning that is given a specific goal (i.e., target variable) for grouping data and allows to allocate observations to various pre-determined segments or class labels. Classification can be used to automatically determine the class of unlabelled or new data. The discovered profiles are labelled objects that can be used to train a classification model and classify new Dutch website visitors interested in Master studies into one of the six behavioural profiles. As a result, a new visitor can readily be (re)targeted according to the characteristics of a particular profile.

Micro-behaviours, and in some cases macro-behaviours, are considerably different across groups and between profiles which allow to tailor marketing campaigns. Hence, future research could focus on studying the differences in micro-behaviours between various countries, study programmes, traffic source, device type or any other meaningful attribute. Moreover, studying specific micro-behaviours of profiles allows to identify what customer attributes influences the conversion rate on Osiris Application Submitted. Furthermore, the Interested-HP and MA-P profiles are influenced by Questions via Web Form which could indicate that some information was not available on the website. Hence, a visitor asked a question which is stored in the form of text and may be voluminous and valuable enough for gaining knowledge about prospects. *text mining* can be considered to analyse the unstructured text and discover what kind of information is asked in order to improve website content and marketing efforts. Additionally, the various hypothesis proposed in this study could be statistically tested. Moreover, future research could test and refine the proposed framework or model by conducting *case studies* with different datasets or utilize them to categorize prior research to find possible research gaps.

This research is limited by the volume, veracity, and variety of the dataset. Volume refers to the data size and veracity to the validity (i.e., accuracy) of the data for its intended use. Therefore, confirmatory research can be conducted using the same methods as in this paper on a dataset of UT website visitors interested in Master studies over a longer period of time. Additionally, a comparison of behavioural profiles of this study with data collected over a longer period might provide insights whether the behaviour among website visitors shift over time. The latter enables the UT M&C department to determine a suitable timeframe to update the profiles, data mining models, and marketing activities. Additionally, the study is limited by the search terms used, journals included, and time period of the papers published. The discovered profiles might not be generalizable. However, the framework and model can be utilized by other organisations and educational institutions to improve data-driven segmentation, User Profiling, and marketing efforts. In addition, confirmatory research can be done by other institutions of higher education to provide evidence of the degree of generalizability of the results. Moreover, the behavioural attributes available in the dataset determine the accuracy of the findings in this paper. Therefore, this study is limited by the *variety* of raw data sources available in addition to the limitations in *volume* and *veracity*.

6. ACKNOWLEDGEMENTS

I would like to thank my family for their support throughout my years of studying. Moreover, my thanks go out to Dr. Ethymios Constantinides, Dr. Sjoerd de Vries, Robert Muster MSc, and Floris Metzner MSc for the opportunity to conduct research on the exciting topic of Behavioural Targeting, their valuable feedback, and patience throughout the Master's Thesis project.

7. REFERENCES

Abbas, A. (2008). Comparisons between data clustering algorithms. *International Arabic Journal of Information technology, 5*(3), 320-325.

Alamuri, M., Surampudi, B. R., & Negi, A. (2014). A survey of distance/similarity measures for categorical data. *Neural Networks International Joint Conference*, 1907-1914. IEEE

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

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.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing, 79*, 3-15. Retrieved from https://arxiv.org/pdf/1312.4722.pdf

Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(4), 340-350. Retrieved from https://doi.org/10.1002/widm.1062

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. Retrieved from https://ieeexplore.ieee.org/document/6949041/

Boerman, S., Kruikemeier, S., & Borgesius, F. (2017). Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising*, *0*(0), 1-14. doi:10.1080/00913367.2017.1339368

Borgesius, F. J. Z. (2016). Singling out people without knowing their names–behavioural targeting, pseudonymous data, and the new data protection regulation. *Computer Law & Security Review*, *32*(2), 256-271.

Bhatnagar, A. and Papatla, P. (2001). Identifying locations for targeted advertisements on the Internet. *International Journal of Electronic Commerce*, *5*(3), 23–44. Retrieved from http://web.b.ebscohost.com.ezproxy2.utwente.nl/ehost/pdfviewer/pdfviewer?vid=1&sid=bb323734-360e-4f9b-8836-05f03de55bd3%40sessionmgr120

Barber, M., Donnelly, K., Rizvi, S., & Summers, L. (2013). An avalanche is coming: Higher education and the revolution ahead. *Institute for Public Policy Research, 11*. Retrieved from http://med.stanford.edu/smili/support/FINAL%20Avalanche%20Paper%20110313%20(2).pdf

Blackboard. (2014). Four Leading Strategies To Identify, Attract, Engage, and Enroll the Right Students, 1–7. Retrieved from https://bbss.blackboard.com/wp-content/uploads/sites/27/2018/07/four-strategies-whitepaper.pdf

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity Measures for Categorical Data: A Comparative Evaluation. *Proceedings of the 2008 SIAM Conference on Data Mining*, 243-254.

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

Bose, I., & Mahapatra, R. K. (2001). Business data mining – a machine learning perspective. Information & Management, 39(3), 211-225.

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. (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., & 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

Chen, J., & Stallaert, J. (2014). An economic analysis of online advertising using behavioral targeting. *MIS Quarterly, 38*(2), 429-449. Retrieved from http://web.b.ebscohost.com.ezproxy2.utwente.nl/ehost/pdfviewer/pdfviewer?vid=1&sid=f02f3af5-568e-4f8d-b064-1e0e38e1e3d 2%40sessionmgr101

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

Cufoglu, A. (2014). User profiling – A short review. *International Journal of Computer Applications, 108*(3). Retrieved from https://pdfs.semanticscholar.org/eecb/f9358916a8e7db20511c611eaceaac554417.pdf

Deane, J., & Meuer, T., Teets, J. (2011). A longitudinal analysis of web surf history to maximise the effectiveness of behavioural targeting techniques. *Electronic Marketing and Retailing, 4*(2), 117-128. Retrieved from https://www.inderscienceonline.com/doi/abs/10.1504/IJEMR.2011 .043037

Demchenko, Y., Grosso, P., De Laat, C., Membrey, P., (2013). Addressing big data issues in Scientific Data Infrastructure. *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS 2013)*: San Diego, CA, p. 48-55.

Diapouli, M., Kapetanakis, S., Petridis, M., & Evans, R. (2017). Behavioural Analytics using Process Mining in On-line Advertising. *Proceedings of the ICCBR 2017 Workshops, 20*(28), 147-156. Retrieved from http://ceur-ws.org/Vol-2028/paper14.pdf

Dolnicar, S. (2008). Market segmentation in tourism. *Tourism management, analysis, behaviour and strategy,* 129-150. Retrieved from https://pdfs.semanticscholar.org/d7d0/1f681371015892e18f c7f68ea9a1dbd878bd.pdf

Dolnicar, S. (2003). Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research, 2003, 11*(2), 5-12.

Dolnicar, S., & Lazarevski, K. (2009). Methodological reasons for the theory/practice divide in market segmentation. *Journal of marketing management, 25*(3-4), 357-373.

Dolnicar, S. (2004). Beyond "commonsense Segmentation"- A systematics of segmentation approaches in Tourism. Journal of Travel Research, 42(3), 244-250.

Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991-16005. Doi: 10.1109/ACCESS.2017.2654247

European Commission (2018). General Data Protection Regulation https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en#documents

Eirinaki, M., & Vazirgiannis, M. (2003) . Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT), 3*(1), 1-17. https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1005&context=computer_eng_pub

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

Fayyad, U., Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, *17*(3), 37-54. doi: https://doi.org/10.1609/aimag.v17i3.1230

Gartner (2016). *Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage*. Retrieved from https://www.gartner.com/newsroom/id/3412017

Goldfarb, A., & Tucker, C. E. (2011). Privacy regulation and online advertising. *Management science*, *57*(1), 57-71. Doi: https://doi-org.ezproxy2.utwente.nl/10.1287/mnsc.1100.1246

Goyat, S. (2011). The basis of market segmentation: a critical review of literature. *European Journal of Business Management*, *3*(9), 45-54. Retrieved from https://www.iiste.org/Journals/index.php/EJBM/article/viewFile/647/540

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, *25*(5), 354-366. https://doi.org/10.1016/S0306-4379(00)00022-3

Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. *AMC Sigmod Record*, *27*(2), 73-84. Retrieved from https://dl-acm-org.ezproxy2.utwente.nl/citation.cfm?doid=276305.276312

Gutwirth, S., Leenes, R., De Hert, P., & Poullet, Y. (2012). *European data protection: in good health?* Springer Science & Business Media. Retrieved from https://www.springer.com/gp/book/9789400729025

Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis Concepts and Methods. *Data Mining*, 443-495. https://doi.org/10.1016/B978-0-12-381479-1.00010-1

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

Hemsley-Brown, J., & Oplatka, I. (2006). Universities in a competitive global marketplace A systematic review of the literature on higher education marketing. *International Journal of Public Sector Management*, *19*(4), 316-338. doi:10.1108/09513550610669176

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, *40*(16), 6491-6507.

Hosseini, M., & Shabani, M. (2015). New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, *3*(3), 110-121. Retrieved from https://link.springer.com/article/10.1057/jma.2015.10

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304. https://doi.org/10.1023/A:1009769707641

Husson, F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2018). Package 'FactoMineR'. [Online] Available: http://mirror.its.sfu.ca/mirror/CRAN/web/packages/FactoMineR/ FactoMineR.pdf [Accessed: 28 June 2018]

Interactive Advertising Board (IAB),The-economic-value-of-data-driven-advertising (2017). [Online] Available: https://www.iabeurope.eu/policy/the-economic-value-of-data-driven-advertising [Accessed: 24 January 2018]

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

Jordan, M. I., & Mitchell, T.M. (2015). Machine learning: trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kailing, K., Kriegel, H. P., Kroeger, P., & Wanka, S. (2003). Ranking interesting

subspaces for clustering high dimensional data. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 241-252). Springer, Berlin, Heidelberg.
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.
Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, 32(8), 68-75.
Khosrow-pour, M. (2005). Encyclopedia of Information Science and Technology. IGI Global. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=3Z6NC01PsLcC&oi=fnd&pg=PR51&dq=khosrow+pour+2009&ots=pWopEY6S6M&sig=Eq2seHZFV0tAzEdgPQ2EwgVownc#v=onepage&q=user%20profiling&f=false
Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks*, *37*(9), 52-65. doi: https://doi.org/10.1016/j.neunet.2012.09.018
Kotler, P. (2000). Marketing Management: The Millennium Edition. *Marketing Management, 23*(6), 188-193.
Kuo, R. J., Ho, L. M., & Hu, C.M. (2002). Cluster analysis in industrial market segmentation through artificial neural networks. *Computers & Industrial Engineering*, *42*(2-4), 391-399.
Kusumawati, A., Yanamandram, V. K., & Perera, N. (2010). University marketing and consumer behaviour concerns: the shifting preference of university selection criteria in Indonesia. *Asian Studies Association of Australia 18th Biennial Conference*, 1-16. Retrieved from http://ro.uow.edu.au/chsd/33/
Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.
Lambrecht, A., & Tucker, C. (2013). When Does Retargeting Work? Information Specificity in Online Advertising. *Journal of Marketing Research, 50*(5), 561–576. doi: http://dx.doi.org/10.1509/jmr.11.0503
Lourenco, F., Lobo, V., & Bacao, F. (2004). Binary-based similarity measures for categorical data and their application in Self-Organizing Maps.
Lewis, E. (1903). Advertising department: catch-line and argument. The BookKeeper, 15, 124–128.
Liu, Y., Kiang, M., & Brusco, M. (2012). A unified framework for market segmentation and its applications. *Expert Systems with Applications, 39*(11), 10292-10302. doi: https://doi.org/10.1016/j.eswa.2012.02.161
Lu, X., Zhao, X., & Xue, L. (2016). Is combining Contextual and Behavioral Targeting Strategies Effective in Online Advertising? *AMC Transactions on Management Information Systems, 7*(1), 1-20. Doi:10.1145/2883816
Romdhane, L. B., Fadhel, N., & Ayeb, B. (2010). An efficient approach for building customer profiles from business data. *Expert Systems with Applications, 37*(2), 1573-1585. Doi: https://doi.org/ 10.1016/j.eswa.2009.06.050
Rika, N., Roze, J., & Sennikova, I. (2016). Factors affecting the choice of higher education institutions by prospective students in latvia. *CBU International Conference in innovation and education, 4*(0), 422-430. doi : http://dx.doi.org/10.12955/cbup.v4.790
Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *AMC transactions on knowledge discovery from data (TKDD), 12*(2),1-68. doi: https://doi.org/10.1145/3132088
Park, H.S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications, 36*(2), 3336-3341. doi:10.1016/j.eswa.2008.01.039
Prasad, Y.L. (2016). *Big data analytics made easy*. USA: Notion Press.

Preeti, P., Kalia, P., & Rani, M. (2016). A new hybrid clustering approach for web mining. *International Journal of Engineering Applied Sciences and Technology, 1*(6), 92-94. Retrieved from www.ijeast.com/papers/92-94,Tesma106,IJEAST.pdf

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

Punj, G., & Stewart, D. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research,* 134-148.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1*(14), 281-297.

Malhotra, N. K. (2004) *Marketing research: an applied orientation, 4th edition,* Prentice-Hall International, London.

Maringe, F. (2006). University and course choice: Implications for positioning, recruitment and marketing. *International Journal of Educational Management, 20*(6), 466-479. doi 10.1108/09513540610683711

Mazanec, J. A., & Strasser, H. (2000). A nonparametric approach to perceptions-based market segmentation: Foundations, (1). Springer

McAfee, A., Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), p. 60-68.

Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate behavioural research, 16*(3), 379-407.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179.

Mota, D., Grilo, A., & Farias, M. (2016). A game-theoretic approach to digital marketing and lead generation for duopoly markets. *Proceedings of the International Conference on Industrial Engineering and Operations Management, 8*(10), 133-142. Retrieved from: http://ieomsociety.org/ieom_2016/pdfs/47.pdf

Muller, H., & Hamm, U. (2014). Stability of market segmentation with cluster analysis – a methodological approach, *Food Quality and preference, 34*(4), 70-78. Doi: http://dx.doi.org/10.1016/j.foodqual.2013.12.004

Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering, 109*, 233-252. doi: http://dx.doi.org/10.1016/j.cie.2017.04.017

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications, 36*(2), 2592-2602. doi: https://doi.org/10.1016/j.eswa.2008.02.021

Ordonez, C. (2003). Clustering binary data streams with K-means. *Data mining and knowledge discovery, 0*(0), 12-19.

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.

Rindfleish, J.M. (2003). Segment profiling: reducing risk in higher education management. *Journal of Higher Education Policy and Management, 25*(2), 147-59.

Saia, R., Boratto, L., Carta, S., & Fenu, G. (2016) Binary sieves: toward a semantic approach to user segmentation for behavioural targeting. *Future Generation Computer Systems, 64*, 186-197. https://www.researchgate.net/profile/Roberto_Saia/publication/301215968_Binary_Sieves_Toward_a_Semantic_Approach_to_User_Segmentation_for_Behavioral_Targeting/links/59de1c4daca272204c2c7d9d/Binary-Sieves-Toward-a-Semantic-Approach-to-User-Segmentation-for-Behavioral-Targeting.pdf

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

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2), 34-38. Doi 10.1109/TITB.2012.2223823

Schiaffino, S., & Amandi, A. (2009). Intelligent User Profiling. *Artificial Intelligence: an International Perspective*, 193- 216. Springer.

Shaw, M., Subramaniam, C., Woo Tan, G., & Welge, M. (2001). Knowledge management and data mining for marketing. Decision Support Systems, 31(2), 127-137. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.1196&rep=rep1&type=pdf

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

Srimani, P. K., & Srinivas, A. (2011). Behavioural Targeting - Consumer Tracking. *American Institute of Physics, 1414*(1), 56-60. doi:10.1063/1.3669931

Signal, H., Kohli, S., & Sharma, A.K. (2014). Web Analytics: State of the art & literature assessment. *5th international conference-confluence the next generation of information technology summit*, 24-29. Doi 10.1109/confluence.2014.6949041

Summers, C., Smith, R., & Reczek, R. (2016). An Audience of One: Behaviorally Targeted Ads as Implied Social Labels. *Journal of Consumer Research*, *43*(1), 156-178. doi: 10.1093/jcr/ucw012

Tamasauskas, D., Sakalauskas, V., & Kriksciuniene, D. (2012). Evaluation framework of hierarchical clustering methods for binary data. *Hybrid Intelligent Systems* (HIS), 2012 12th International Conference, 421-426. Doi: 10.1109/ICHPCA.2014.7045336

Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data Mining Techniques in CRM: inside customer segmentation*. John Wiley & Sons. Retrieved from https://pdfs.semanticscholar.org/e4ca/f946c219d9afd433dcad02679ba346e8d5ce.pdf

Lorenco, F., Lobo, V., & Bacao, F. (2004). Binary-based similarity measures for categorical data and their application in Self-Organizing maps, 1-18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.9166&rep=rep1&type=pdf

Turing, A. M., (1950). Computing machinery and intelligence. *Mind 59*(236), 433–460.

UTwente Educational Facts and Figures. Retrieved January 8, 2018, from https://www.utwente.nl/en/facts-and-figures/education/

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

Vázquez, S., Muñoz-García, Ó., Campanella, I., Poch, M., Fisas, B., Bel, N., & Andreu, G. (2014). A classification of user-generated content into consumer decision journey stages. *Neural Networks, 58*(9), 68-81. Doi: http://dx.doi.org/10.1016/j.neunet.2014.05.026 0893-6080

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly.

Walters, M., & Bekker, J. (2017). Customer super-profiling demonstrator to enable efficient targeting in marketing campaigns. *South African Journal of Industrial engineering, 28*(3), 113-127. Retrieved from sajie.journals.ac.za/pub/article/download/1846/807

Wijaya, B. S. (2012). The Development of Hierarchy of Effects Model in Advertising. *International Research Journal of Business Studies*, *5*(1), 73–85. Retrieved from http://management-update.org/uploads/dokumen/5-1-e.pdf

Wolfswinkel, J.F., Furtmueller, E., & Wilderom, C.P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems, 22*(1), 45-55.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How much can behavioral targeting help online advertising? *In Proceedings of the 18th international conference on World wide web.* pp. 261-270. Retrieved from http://citeseerx.ist. psu.edu/viewdoc/download?doi=10.1.1.215.1473&rep=rep1&type=pdf

Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation. *Harvard business review, 84*(2), 122-139 retrieved from http://viewpointlearning.com/wpcontent/uploads /2011/04/ segmentation_0206.pdf

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

Zhao, Y., & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *International conference on information and knowledge management*, 515-542.

APPENDIXES

Appendix A

Literature Search Strategy:

A systematic literature review is conducted with the support of the methods as described in Wolfswinkel, Furtmueller and Wilderom (2013) and Webster and Watson (2002). A computer search has been done by accessing different scientific search engines such as Scopus (world's leading database), Web of Science, and Google Scholar. Only articles published in academic journals and conference proceedings have been used for the literature review. After initial searches with relevant keywords and search strings, the articles were sorted by relevance and the abstracts were read. Additionally, the relevant articles were further filtered by the amount of citations to find high-impact articles or journals. First, selection is done by evaluating the title and abstract. Next, relevant papers were selected by comparing abstracts, citations, and finally by reading the full text. Among others the following keywords were used to create search stings and find relevant literature: (1) Behavioural/Behavioral Targeting, (2) Behavioural/Behavioral advertising/advertizing, (3) Machine learning, (4) customer segmentation, (5) customer journey, (6) (online) consumer behaviour/behavior, (7) digital marketing, (8) Higher Education, (9) Data Mining, (10) online user tracking, (11) user profiling. The table below presents a brief overview of the search strings that were used, and continuously refined, to collect literature.

Appendix B

Overview of DM and KD process models

Model/Ref	Fayyad et al. (1996) (KDD)	Canbena et al. (1998)	Anand & Buchner (1998)	CRISP-DM (Shearer, 2000)	Cios et al. (2000)	SEMMA (Sas Institute, 2005)
Area	Academic	Industrial	Academic	Industrial	Academic	Industrial
No. steps	9	5	8	6	6	5
Steps	Developing and Understanding the application domain	Business objectives	Human resource identification Problem specification	Business understanding	Understanding the problem domain	
	Selection of target data	Data preparation	Data prospecting Domain knowledge elicitation	Data understanding	Understanding the data	Sample Explore
	Data pre-processing		Methodology Identification	Data preparation	Preparation of data	Modify
	Data Transformation		Data pre-processing			
	Choosing the DM task					
	Choosing the DM algorithm					
	DM	DM	Pattern discovery	Modelling	DM	Model
	Pattern Interpretation/Evaluation	Domain knowledge Elicitation	Knowledge Post-pre-processing	Evaluation	Evaluation of discovered knowledge	Assessment
	Consolidating Discovered knowledge	Assimilation of knowledge		Deployment	Using the discovered knowledge	

Note. (Adapted from Kurgan, & Musilek, 2006; Fayyad, Shapiro, & Smyth, 1996; Shafique & Qaiser, 2014; Larose, 2014).

Appendix C

Overview of User Profiling 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 D

Hybrid Profiling methods by Poo et al. (2003).

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**

Appendix E

Overview of User Profiling 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 F

*Performance of ten hierarchical clustering methods based on various symmetric distance measures (*Tamasauskas et al., 2012).

TABLE 7. RESULTS OF SYMETRIC DISTANCE MEASUREMENTS

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

Appendix G

Categorization of clustering algorithms and big data properties (Fahad et al., 2014)

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

Appendix H

Complete Distribution of Study Programmes for All Master Visitors

Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5			Cluster 6		
Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%
SET	67	15.9	HS	17	12.4	EE	30	9.8	SE	108	12.8	CEM	21	18.8	ME	195	10.9
ME	48	11.4	SE	14	10.2	BA	28	9.2	GISEO	84	9.9	SET	19	17.0	CPS	170	9.5
CEM	47	11.2	CEM	13	9.5	CPS	27	8.8	EEM	83	9.8	ME	17	15.2	EE	170	9.5
IEM	44	10.5	CE	12	8.8	SET	27	8.8	ME	69	8.1	CME	11	9.8	SET	170	9.5
CME	30	7.1	CME	9	6.6	ME	26	8.5	CEM	54	6.4	BA	9	8.0	CEM	143	8.0

EE	26	6.2	BA	8	5.8	CEM	20	6.5	AM	42	5.0	IDE	9	8.0	CE	103	5.8
CE	25	5.9	CPS	8	5.8	BE	19	6.2	IEM	42	5.0	IEM	9	8.0	CME	100	5.6
CPS	25	5.9	IDE	8	5.8	IEM	16	5.2	BA	41	4.8	EE	8	7.1	BA	95	5.3
BA	24	5.7	IEM	8	5.8	P	15	4.9	CE	41	4.8	CE	7	6.3	ES	95	5.3
BIT	24	5.7	BIT	7	5.1	ES	14	4.6	SET	37	4.4	HS	7	6.3	BE	90	5.0
AM	21	5.0	ME	7	5.1	HS	14	4.6	HS	34	4.0	N	6	5.4	BIT	90	5.0
SC	21	5.0	AM	5	3.6	CE	13	4.2	CME	33	3.9	SC	6	5.4	PSTS	81	4.5
HS	20	4.8	BE	5	3.6	BIT	12	3.9	CPS	31	3.7	CPS	5	4.5	CS	73	4.1
IDE	20	4.8	CS	5	3.6	CME	12	3.9	EE	26	3.1	ES	5	4.5	IEM	69	3.9
CS	19	4.5	ES	5	3.6	IDE	11	3.6	CS	24	2.8	BE	4	3.6	EST	58	3.2
ES	15	3.6	EST	4	2.9	N	11	3.6	N	21	2.5	BIT	4	3.6	N	57	3.2
EEM	14	3.3	SET	4	2.9	SC	11	3.6	BIT	19	2.2	CS	4	3.6	AM	54	3.0
BE	13	3.2	SC	3	2.2	CS	10	3.3	EST	19	2.2	EST	4	3.6	SC	44	2.5
IST	13	3.1	EE	2	1.5	AM	8	2.6	IDE	18	2.1	IT	4	3.6	PA	40	2.5
AP	11	2.6	EEM	2	1.5	EST	8	2.6	IT	17	2.0	EEM	3	2.7	IDE	38	2.1
EST	10	2.4	N	2	1.5	IT	8	2.6	AP	16	1.9	P	3	2.7	PSTS	37	2.1
N	10	2.4	P	2	1.5	EUR	7	2.3	BE	15	1.8	AM	2	1.8	HS	31	1.7
PSTS	10	2.4	TM	2	1.5	PSTS	6	2.0	IST	13	1.5	AP	2	1.8	AP	30	1.7
IT	9	2.1	AP	1	0.7	PA	6	2.0	P	13	1.5	PA	2	1.8	IT	28	1.6
P	9	2.1	ES	1	0.7	EEM	4	1.3	PA	11	1.3	EUR	1	0.9	EUR	25	1.4
PA	7	1.7	IST	1	0.7	IST	4	1.3	TM	11	1.3	IST	1	0.9	IST	24	1.3
SE	7	1.7	PA	1	0.7	AP	2	0.7	WT	11	1.3	PSTS	1	0.9	TM	1	0.1
ES	4	1.0				SE	1	0.3	PSTS	10	1.2	SE	1	0.9			
GISEO	2	0.5							ES	7	0.8	WT	1	0.9			
TM	2	0.5							SC	5	0.6						
									EUR	2	0.2						

Appendix I

Complete Distribution of Study Programmes for Dutch Master Visitors in each cluster

Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5			Cluster 6		
Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%	Studies	N	%
HS	5	20.8	BA	8	15.1	CS	5	29.4	CS	17	11.3	HS	16	51.6	HS	16	11.7
BA	4	16.7	HS	6	11.3	CEM	3	17.6	P	13	8.7	BE	2	6.5	BA	13	9.5
AM	3	12.5	EE	5	9.4	CME	3	17.6	SET	12	8.0	COMPS	2	6.5	EEM	13	9.5
CEM	3	12.5	SET	5	9.4	BA	2	11.8	ME	11	7.3	IEM	2	6.5	SE	13	9.5
EE	3	12.5	CE	4	7.5	HS	2	11.8	BA	10	6.7	TM	2	6.5	CEM	11	8.0
IDE	3	12.5	P	4	7.5	IT	2	11.8	BIT	10	6.7	AM	1	3.2	EST	9	6.6
COMPS	2	8.3	CS	3	5.7	P	2	11.8	CEM	10	6.7	AP	1	3.2	IT	8	5.8
EST	2	8.3	ME	3	5.7	AM	1	5.9	CS	9	6.0	BA	1	3.2	SET	8	5.8
IEM	2	8.3	PSTS	3	5.7	CE	1	5.9	BE	8	5.3	BIT	1	3.2	BIT	7	5.1
ME	2	8.3	CEM	2	3.8	EE	1	5.9	EST	8	5.3	CE	1	3.2	CS	7	5.1
SC	2	8.3	IDE	2	3.8	EEM	1	5.9	CE	7	4.7	CEM	1	3.2	GISEO	7	5.1
BE	1	4.2	N	2	3.8	IST	1	5.9	PH	7	4.7	CME	1	3.2	ME	7	5.1
P	1	4.2	PA	2	3.8	ME	1	5.9	AP	6	4.0	EST	1	3.2	BE	6	4.4
PA	1	4.2	BE	1	1.9	PSTS	1	5.9	EL	6	4.0	IDE	1	3.2	CE	6	4.4
SE	1	4.2	BIT	1	1.9	SET	1	5.9	HS	6	4.0	IT	1	3.2	IEM	6	4.4
SET	1	4.2	CME	1	1.9	SC	1	5.9	ES	5	3.3	N	1	3.2	TM	6	4.4
			EEM	1	1.9	TM	1	5.9	TM	5	3.3	SE	1	3.2	CS	5	3.6

ES	1	1.9	CME	4	2.7	SC	1	3.2	CME	5	3.6
IEM	1	1.9	SC	4	2.7				AP	4	2.9
IT	1	1.9	BIT	3	2.0				ES	4	2.9
IST	1	1.9	PA	3	2.0				IDE	4	2.9
SC	1	1.9	IN	2	1.3				AM	3	2.2
			IDM	2	1.3				EE	3	2.2
			IT	2	1.3				PSTS	3	2.2
			N	2	1.3				IST	2	1.5
			EEM	1	0.7				P	2	1.5
			SP	1	0.7				PA	2	1.5
									ES	1	0.7
									N	1	0.7
									SC	1	0.7
									WT	1	0.7

Appendix J

Methods, packages, and functions used in R

Stage 1	Method	Package	Function	Version
1	Calculating Hamming Distance	EnsCat	hammingD()	1.1
2	Distance object	stats	as.dist()	3.5.1
2	Hierarchical Clustering (complete/hamming)	stats	hclust()	3.5.1
3	Plot of dendrogram	stats	plot()	3.5.1
4	Plot of best partition	Jlutils	best.cutree()	1.14.0
Stage 2				
5	K-modes	KlaR	kmodes()	0.6-14
6	Descriptive Statistics	pastecs	stat.desc()	3.1.21
7	Write factor $cluster to Excel	openxlsx	write.xlsx()	4.1.0
8	Data Partition for Cross-Validation	Caret	createDataPartition()	6.0-80
9	Silhouette object	stats	silhouette()	3.5.1
10	Plot Silhouette	factoextra	fviz_silhouette()	1.0.6