

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Detecting different types of customer behaviour using customer journey analytics

Jorrit Jespers

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR:

dr. Gert JANSSENSWILLEN



oww.uhasselt.he

universiteit Hasselt Campus Hasselt: Aartelarenlaan 42 | 3500 Hasselt Campus Diepenbeek: Igoralaan Gebouw D | 3590 Diepenbee $\frac{2020}{2021}$



Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Detecting different types of customer behaviour using customer journey analytics

Jorrit Jespers

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR:

dr. Gert JANSSENSWILLEN

Covid disclaimer

This master thesis was written during the COVID-19 crisis in 2020-2021. This global health crisis might have had an impact on the (writing) process, the research activities and the research results that are at the basis of this thesis.

Detecting different types of customer behaviour using customer journey analytics

Jorrit Jespers^{1,2}

Promotor: Gert Janssenswillen²

Business and information systems engineering
UHasselt - Hasselt University, Faculty of Business Economics, Agoralaan, 3590
Diepenbeek, Belgium

Abstract. With companies always looking for more and better ways to leverage data and use it to their advantage, the analysis of customer journeys can add a valuable alternative. The field of customer journey analytics experiences a shift towards data-driven and analytical methods, supported by the increasing wealth of available data. This resulted in the distinction between an ideal customer journey and an actual customer journey, where the latter is based on behavioural customer data. To this day, there is no generally accepted definition of 'a customer journey'. Therefore, we propose a formal definition of the concept in the light of the customer journey analytics field. To further demonstrate the value of a more data-driven approach, the goal is to make the representative journey, which is a high-level representation of the actual customer journey, deliver more usable information. This is done by creating different clusters where for each cluster process mining techniques are used to extract the actionable differences between them. We conclude that it is possible to create clusters based on customer journey data, whereby differences are discovered between the behaviour that these clusters produce. These differences can be used with the goal to support choices with regard to marketing tactics. Besides those tactics, the gained insights can also support the classification of new customers and be the basis to start predicting the next touchpoint of customers.

Keywords: Customer journey analytics - Trace clustering - Behavioural clustering - Actual customer journey - Process mining

1 Introduction

The use of customer journeys (CJs) originated in several industries, therefore the roots can be found in practice. For example, in marketing, CJs are used to become aware of a customer's decision-making process, while in service design they use CJs to outline the service. Finally, in service management, CJs are used to strengthen the experience throughout the entire service process [17].

2 Jorrit Jespers

Only quite recently the subject started to appear in Scientific literature, where the first studies on CJs intended to merge the various applications in practice into an integrated research domain [17]. Accordingly, there was a strong focus on the differences in how these domains define a CJ and how to create a formal definition [24, 46].

As a result of the rapid development of various domains, including process mining and data mining, together with the ever-growing amounts of data, a data-driven vision emerged in the domain of CJs. The rationale behind this new discipline is the difference between a planned/ideal CJ and the real/actual CJ. In other words, until recently, no distinction was made between these different perspectives in literature. A CJ was considered from the company's perspective (planned CJ), but they did not discuss what the customers really experienced (actual CJ). The critique highlighted that it is very unlikely that the company's perspective and the customer's experience coincide. A second point was that every individual journey can look different. The term used to combine the research that tries to close the gap between planned and actual CJ is coined 'customer journey analytics' (CJA) [24]. In order to further close that gap and especially to increase the value of the actual CJ, the gained information needs to be more fine-grained. Generating a representative CJ using sequential customer data is already proven [4]. The representative CJ compromises a single high-level map of the actual CJ, which is too general. A more detailed view on different journeys will deliver more valuable insights.

With the emergence of CJA in the novel domain of CJs in literature there is still a need to define a CJ within the context of CJA. Therefore, a formal definition of the CJ will be proposed. This definition is also the foundation for the other concepts presented in this research paper. In order to reach a more fine-grained level of information using the CJ, different techniques can be used. Within this paper the goal is to broaden the view on the CJ by creating different segments of customers. This will be done by using clustering methods like agglomerative hierarchical clustering and k-medoids clustering. Once the clusters are identified the goal is to differentiate between these segments. By using these differences, one can make better-informed decisions on their marketing tactics.

The paper is structured as follows. Section 2 describes the related work upon which the paper will build. A more detailed overview of the problem statement will be given in section 3. In section 4 we propose a formal definition of the CJ. Section 5 highlights the methodology of this paper. In section 6 the results of the clustering approach are discussed and the clusters are analysed, followed by a discussion in section 7. Section 8 describes the overall conclusions of this work.

2 Related work

Customer journey analytics sits in between different domains. Further more, CJA is closely intertwined with the more general topic of CJs. Although there is no generally accepted definition available for what a CJ exactly entails [17], in the literature review of Folstad and Kvale [17] the main aspects and possible

definitions were explained. The first aspect in which there is variation within the definition is how the start and end should be defined. One could say that there should be a well-defined start and endpoint, while others will argue that the CJ can have an open ending. Second, the stages or steps within a CJ are discussed together with the scope. A first possibility is to take the CJ in the narrow sense of the word, and only focus on the service. On the other hand, it is also possible to look at what happens before and after the service [45]. Although the stages prepurchase, purchase and post-purchase are used by multiple authors this is not always the case [35, 46]. Some studies still opt for stages that specifically depend on the scope of their study [38, 39]. Lastly, defining what a touchpoint means within the CJ is important. Touchpoints are used in two different ways, first being the interaction/communication between the customer and the service [39, 49]. Alternatively, a touchpoint can be seen as the location or channel mediating the interaction or communication [13]. This discussion resulted from the different ways of using CJs in practice.

During the last few years a new dimension was added to the definition. It is no longer 'the customer journey', but a distinction is made between a planned and actual CJ. Before, literature was mostly dealing with the planned CJ [4, 24]. A planned CJ is a specification of the CJ the company thinks their customer will go through [26]. On the other hand, the actual CJ is the representation of the steps the client actually goes through [24, 49]. It tries to shape what the CJ looks like in the real world. This can be done by collecting data through surveys or field studies [39].

In the most recent years the first initiatives were taken towards a more datadriven approach, that uses customer data and event data stored in company data warehouses [7, 23]. The biggest benefit of these new approaches is that there is no longer a need for intensive and costly data gathering methods like interviews or surveys in order to discover the actual CJ [35]. The opportunity to use the already available data from the company on its customers will pave the way towards a more objective view on what the customer really goes through.

In order to get that more objective view on the CJ, a first step was to create customer journey maps (CJM) based on sequential customer data. A CJM is a tool used to visualise the service the customer goes through. In order to be able to create these CJMs based on data, different methods were presented, which mostly stem from the process mining field, more specifically form process discovery techniques [3–7, 49]. Markov-chains are also proposed as an alternative method that can be used to create a mapping of the CJ [1, 15, 25].

A next step is to start differentiating the CJ of different customers. Not every customer goes through the same CJ. Although the techniques in the previous paragraph are capable of discovering the CJ by using data, these discovery algorithms build some sort of a representative journey, i.e., they return one generic journey that is covering many different individual journeys [4]. To get a more detailed and better understanding of all customers, it can be valuable to create different segments. Because in case only a generic CJ is available, there is a loss of information about all individual customers. A first approach, to counter

4 Jorrit Jespers

this information loss, remains within the domain of process mining and uses techniques to cluster traces, where each trace represent a journey of a specific customer [19, 48, 49]. A second method consists of transforming the sequential data, from which the CJ stems, into a dataset using features. Based on these features the techniques identified in the field of data mining/clustering and customer segmentation can be used [27, 40, 42, 53].

Once different clusters are identified, they can deliver more insights to the business. From a marketing perspective, having a generic CJ can be an eye opener, showing the difference between the actual and planned CJ, but adds little actionable insight. This information is too general to create marketing tactics. In an ideal situation each customer is approached individually, unfortunately this is hard to achieve. Therefore, having segments is an ideal compromise to determine marketing tactics and to gain a competitive advantage within each segment [21, 30, 37]. On the other hand, these segments can also be used by the business to classify new customers. While the information about new customers is usually limited, it can be used, together with the identified segments, to perform a classification [12, 51, 56]. The classification algorithm will give back probabilities about how likely it will be for that customer to belong to a certain segment. Based on that information, the business, even at an early stage, can optimize its cluster handling and determine the most appropriate marketing techniques [50].

3 Problem statement

The first step towards a more objective view on the CJ already received considerable attention, by developing different ways to discover the real CJ. Although this is a step in the right direction, there is still room for improvement within the field of CJA. At this moment, the techniques used, return a representative CJ, but this is only very high level, and there is still much generalization going on. Hence, the extent to which marketing decisions can already be supported is rather limited. It is important to extent the real potential of CJA by discovering different segments.

Currently, a CJM delivers insights only to a minimal extend. For example, it can help in redesigning the general process, although not every customer goes through the same process, which is demonstrated by the growing amount of data. Therefore, a need for a more detailed representation of differences between the journeys of customers exists. Probably almost every CJ is unique, but it is practically not feasible to work on an individual level and support decisions based on this. On the other hand, only using one representative CJ is too general as well. A lot of information is lost on less frequent traces. The map gives a representation of the frequent customer paths. It may be that some paths and customers occur less frequently, but that does not make them less valuable to the company. Therefore, creating segments, with each segment having its own representative journey, seems the right compromise.

When using different segments, this can fully unlock the potential to support marketing decisions. If one would have more details about the different groups, the process could be better developed for each group separately. The ultimate goal of marketing is to ensure that each customer brings as much value to the business as possible. This is done by using different marketing techniques. In order to be able to make decisions about these different marketing techniques, there is a need for more detailed information [37]. A step towards this more detailed view can be delivered by adopting the reasoning behind creating segments within the domain of CJA.

4 Defining the customer journey

The absence of a commonly accepted definition of a CJ, partially stems from the fact that the CJ finds its origins in different domains in practice [17]. Most of the papers use the definition of a CJ within their domain, without considering the others. As the practical domains increasingly converge into one domain in literature, i.e., customer journey analytics, the definitions should also be brought together. While some authors mention this need and others have already made an attempt, there is yet no general accepted definition [17, 24]. However there are certain elements that authors agree on, i.e. a CJ [4, 7, 16, 17, 20, 24, 28, 31, 35, 46, 47, 55]:

- represents the purchase process of a customer interacting with a firm over time
- consists of touchpoints between the customer and the firm
- focuses on the perspective of the customer

The paper of Folstad and Kvale [17] clearly identified and illustrated the discrepancies that are still present. They identified five areas where literature has different views on the definition of the CJ. (1) Some papers look at the CJ as a one-time cycle whereas other see it as a continuous process. (2) There is some variation in how the start- and endpoint of a CJ should be defined. (3) There is also disagreement in terms of the amount of different channels a CJ can comprise. A channel is defined as the medium through which a customer and a company interact [43]. (4) Touchpoints are used in two different ways: as the interaction/communication between the customer and the service or alternatively, the channel mediating the interaction or communication [14, 39, 49].(5) The variation in the sectioning of the CJ. These sections are also called steps or stages of the CJ. While some authors use predefined stages (e.g. pre-purchase, purchase and post-purchase) others still opt for stages dependent on the scope of their study [35, 38, 39, 46].

In the next paragraphs, we will build a formal definition of the CJ. To this end, we need to define the touchpoints that make up the journey. Therefore, we define a universe of all possible touchpoints that can occur.

Definition 1. Let $\mathcal{T} = \{t_1, ..., t_n\}$ be the universe of possible touchpoints.

To further define a touchpoint, we split it into different components. The first component is the type. This defines what the goal of the touchpoint is. Lemon and Verhoef [35] identified four categories: brand-owned, partner-owned, customer-owned and social/external owned touchpoints. The brand-owned touchpoints are customer interactions during the experience that are designed and managed by the firm and under the firm's control (e.g. advertising, websites, service). The customer interactions that are jointly designed, managed, or controlled by the firm and one or more of its partners are the partner-owned touchpoints (e.g. multichannel distribution, multivendor loyalty program, marketing agency). The customer-owned touchpoints are customer actions that are part of the overall customer experience, but that the firm, its partners nor others can influence or control (e.g. customer thinking about needs and desire). The social/external owned touchpoints refer to the external factors, like other customers or independent information sources, that have an impact on the customer experience.

Definition 2. Let $C = \{c_1, ..., c_n\}$ be the universe of possible types or categories of the touchpoint.

When looking at the type of touchpoint, there are different parties that can own the touchpoint. This does not define who initiates the touchpoint. Furthermore, every touchpoint has a certain goal, for example a sale, obtaining information, returning goods or making a complaint. This will be defined by the message component.

Definition 3. Let $\mathcal{M} = \{m_1, ..., m_n\}$ be the universe of possible messages or content of the touchpoint.

Every interaction that takes place between the customer and the company happens via a channel. Here, the channel is defined as the medium through which this interaction was made possible (e.g. mail, physical visit, call). Therefore, the universe of all possible channels is defined as \mathcal{V} .

Definition 4. Let $\mathcal{V} = \{v_1, ..., v_n\}$ be the universe of possible channels.

Now that we have defined all the components, one can say that a touchpoint is defined by the combination of a type, a message and a channel.

Definition 5. A touchpoint t_i , where $t_i \in \mathcal{T}$, is a tuple (c, m, v), where $c \in \mathcal{C}$, $m \in \mathcal{M}$ and $v \in \mathcal{V}$

Since a CJ consists of a sequence of touchpoints, we first define the set \mathcal{J} that contains all possible sequences that can be formed with the touchpoints present in \mathcal{T} .

Definition 6. Let $\mathcal{J} = \mathcal{T}^*$ be the set of all finite sequences over \mathcal{T} . A customer journey $\sigma \in \mathcal{J}$ is a sequence of touchpoints σ .

The only thing that is still needed to define a CJ, is making the relation between a customer and a sequence of touchpoints. This is done by defining a CJ as a combination of a certain customer coming form the universe of all customers and a sequence formed by touchpoints.

Definition 7. Let $\mathcal{L} = \{l_1, ..., l_n\}$ be the universe of possible customers.

Definition 8. Let $CJ = (l, \sigma)$ be a customer journey instance where $l \in \mathcal{L}$ refers to customer and where $\sigma \in \mathcal{J}$ refers to a specific customer journey.

Note that using the above definition, the CJ is seen as a continuous process. In theory a sequence can be infinite, but in a dataset the number of combinations will always be finite. Therefore, Definition 6 defines a sequence of touchpoints as finite. However this does not imply that the CJ itself is finite, as there is no such thing as an endpoint. The startpoint will be the first touchpoint the customer has with the company. A touchpoint is defined by its components. By including the channel into the touchpoint, the use of multiple channels is allowed. This is valuable since the environment will be increasingly multi- and omnichannel [2, 34]. Lastly, the sectioning can be dealt with by using a mapping function to assign touchpoints to a specific stage (e.g. pre-purchase, purchase, post-purchase). First all stages are defined by creating the universe of stages.

Definition 9. Let $S = \{s_1, ..., s_n\}$ be the universe of possible stages.

The goal is to map each touchpoint present in the sequence of touchpoints in the CJ on a certain stage. This is done by the mapping function \mathcal{F} .

Definition 10. Let \mathcal{F} be the mapping function, then $\mathcal{F}: t \Rightarrow s$ where $t \in \mathcal{T}$ refers to a specific touchpoint and $s \in \mathcal{S}$ refers to the stage.

The mapping of the touchpoints on the stages does not change the meaning of what a CJ entails. Therefore, this step highly depends on the user. First of all, the stages are defined by the user. These can be pre-defined stages like the pre-purchase, purchase and post-purchase but also stages that depend on the scope of the study. This step is valuable since it can deliver an extra dimension to better understand the journey of your customer but is still optional. If and how this step is filled in, is therefore up to the creator of the CJ.

5 Methodology

5.1 The dataset

The dataset used in this study comes from the platform Kaggle ³. It contains 2 456 414 touchpoint observations about customers in the Netherlands on the purchase/customer journeys from people interacting with a travel agency. The

 $^{^3}$ https://www.kaggle.com/kishlaya
18/customer-purchase-journey-netherlands/version/1

Jorrit Jespers

8

dataset concerns the touchpoints of 9 678 different customers over a period of 17 months, from May 2015 until October 2016. For each customer there are one or multiple purchase journeys, each identified by a unique purchase ID. For every touchpoint a timestamp was registered, together with the touchpoint. There are 20 different touchpoints included in the dataset, shown in Table 1. The touchpoints labelled 1 up to 10 and 12 up to 16 are customer initiated touchpoints, while the touchpoints labelled 18 up to 22 are company initiated. The message column was added manually to the data. This message classification was generated based on the type of touchpoint and holds the information about the content of the touchpoint (see Definition 3). The column initiator was also added to the dataset, which can not be considered as the type (see Definition 2) since this column focuses on who initiates that touchpoint and not who owns that touchpoint. Ideally this was the type of the touchpoint, nevertheless the proposed approach stays applicable.

Table 1. Touchpoints in dataset.

$\overline{\text{Code}}$	Touchpoint	Initiator	Message
1	Accommodations Website	Consumer	Accommodations
2	Accommodations App	${\rm Consumer}$	Accommodations
3	Accommodations Search	${\rm Consumer}$	Accommodations
4	Information / comparison website	${\rm Consumer}$	Information
5	Information / comparison App	${\rm Consumer}$	Information
6	Information / comparison Search	${\rm Consumer}$	Information
7	Tour operator / Travel agent Website Competitor	${\rm Consumer}$	Tour operator
8	Tour operator / Travel agent App competitor	${\rm Consumer}$	Tour operator
9	Tour operator / Travel agent Search Competitor	${\rm Consumer}$	Tour operator
10	Tour operator / Travel agent Website Focus brand	${\rm Consumer}$	Tour operator
12	Tour operator / Travel agent Search Focus brand	${\rm Consumer}$	Tour operator
13	Flight tickets website	${\rm Consumer}$	Flights
14	Flight tickets App	${\rm Consumer}$	Flights
15	Flight ticket Search	${\rm Consumer}$	Flights
16	Generic search	${\rm Consumer}$	Generic
18	Affiliates	Company	Indirect marketing
19	Banner	Company	Direct marketing
20	Email	Company	Direct marketing
21	Prerolls	Company	Direct marketing
22	Retargetting	Company	Direct marketing

More detailed information about the duration of the touchpoint, the device, whether or not the touchpoint is associated with a purchase at a travel agency or competitor and for how long the customer is part of the mobile or fixed panel are also included in the data set. In order to make the approach generically applicable only the Customer ID, the timestamp, touchpoint, message and initiator were used. The rest of the variables are case specific and not available in a generic dataset and therefore left aside.

The type of data was a major challenge within this research. The data that was used is sequential data, which means that the observations run over a period of time. A possible solution to better deal with this time dimension was to use windows, also called time intervals [8, 44]. A second consequence of the type of data are the large differences in the length of CJs. Because the occurrence of a touchpoint is the basis of the observations, and thus did not happen at fixed times. As a result the traces can highly differ in length. When clustering without controlling for the journey length, the resulting clusters will be influenced by loyalty and therefore less reflect the displayed behaviour. Again using windows could be beneficial for the quality of the clusters. Another option is to normalise the calculated distance between traces to control for the length difference. Both methods will be examined during the testing of different scenarios.

5.2 Customer journey clustering

5.2.1 Approach

Dividing customers into groups is a problem that already got a lot of attention in literature. In marketing, one saw that more and more data was being collected and stored about customers. In order to improve marketing efforts, one used this data and applied knowledge discovery and data mining techniques. The purpose of these techniques was segmentation, but also forecasting and developing knowledge-based marketing decision support systems [29]. As Nairn and Berthon described, the main idea of segmentation or clustering is to form groups of similar customers. That group is described by a set of customers who have similar characteristics [41]. These characteristics can be defined based on demography, behaviour, values, and so on, all present in the collected data. The difference between clustering CJs and existing approaches is the type of data. For example, the demographic data or values like sales numbers are used as features. These features can be registered through time, but the data point itself also contains a certain degree of information. On the other hand, the CJ data is a log of touchpoints that happened over time. Only the chain of these touchpoints contains a certain value. So in this case you need to hold together the chain of touchpoints that were logged for a certain customer through a certain time period.

In process mining different approaches to cluster sequential data are already applied. The goal for trace clustering the CJs is to define a notion of similarity or dissimilarity between individual CJs and then cluster the log of all the CJs into k-clusters (for some $k \ge 2$) such that all journeys within a cluster are similar in some sense, and journeys belonging to different clusters are dissimilar [10]. The more traditional trace clustering approaches transform the traces into a vector space [52].

A first approach is called, bag-of-activities. Within this approach a trace is transformed into a vector, where each dimension of the vector corresponds to an activity. The number of dimensions is thus defined by all the activities present in the CJ log. The values in the vector correspond to the frequency count of the

activities in that trace. The similarity is estimated by using standard distance metrics like an Euclidean distance. Terragni and Hassani [49] identified that this approach focuses on the relation between the behaviour of the customer and the activity itself. This can be a drawback since the approach does not capture the dynamics or context of the journey, nor the order of execution is taken into account [10]. A second approach gives the possibility to incorporate more context into the vector which can be done by using an n-gram model instead. For example, when n equals two, the journey will be split up in pairs of two touchpoints. This way the traces are transformed in a vector space. The numbers now represent how many times that gram is present in the CJ [11]. The biggest drawback of this method is that the complexity of the calculations increases with the size of the alphabet, in this case the number of possible touchpoints, and the chosen number for n [33].

The previous two approaches are using a vector-space to determine the similarity. In other words, these approaches are feature driven, whilst these next approaches are trace driven. The hamming distance and edit distance are two methods that rely on the syntax to determine the similarity. The hamming distance is defined as the minimum number of symbols that need to be changed to convert one journey into the other. In order to calculate the Hamming distance, it is required that both journeys have the same length, since this type of distance is not defined for sequences of different lengths [9]. Another drawback is that, according to Bose and van der Aalst, the interleaving of activities is punished too strongly [10]. Therefore, using the Levenshtein edit distance, also called the optimal matching distance, could be the right choice for trace clustering within the context of CJs. The distance is defined as the minimum number of needed operations to transform one sequence into the other, where an edit operation is an insertion, deletion or substitution of an activity. The biggest stumbling block in using this method is determining the cost of the various operations. The major advantage of defining distance in this way, is that the context as well as the order of the trace is preserved [49].

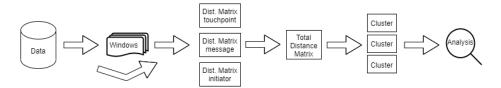


Fig. 1. Overview used approach.

The approach proposed and used in this paper is visualised in Figure 1. The starting approach, directly uses the data for calculating the distance matrices which are also normalised. Then it brings the different matrices together into one distance matrix using a weighted average and applies a clustering algorithm. Although this approach is the starting approach, different scenarios are created

and tested using different settings. The settings that can be changed over the different scenarios are the following: using windows or not, using normalisation or not, the weights to create total distance matrix and the clustering technique. In order to try and control for length differences in journeys, both windows and normalisation were tested. If windows are used these are applied before calculating the distance. The length of the windows used here are quarters. Meaning, all the CJs were split into seven different quarters. The use of the windows ensures that the large differences in length are reduced. This also has the advantage that if there is still a large difference in the number of touchpoints, it is no longer due to loyalty, but is the result of the actual behaviour shown during a certain period of time. In other words, this means that the customer was more active during that particular quarter than the customer for whom fewer touchpoints were observed. A drawback of using windows is that the size of the distance matrix increases significantly by splitting up the CJs. When windows are used, the role of distance normalisation can be diluted. Therefore, we will also look at the scenarios where no normalisation is used and thus only windows. Further explanation on normalisation and the weights is provided in section 5.2.2 and in section 5.2.3 the clustering step is explained. Once the clusters are identified, the log off all CJs can get split up into the different subsets. Subsequently, these subsets can be analysed based on the differences in touchpoints.

5.2.2 Calculating distance

Before the dataset could be used for clustering, the distance between all traces must be calculated. The dataset was in a format where each row contained a touchpoint with the corresponding UserID and time of observation. In order to be able to calculate the distance some pre-processing was necessary to change the format. In the new format one sequence was created per UserID, where the states are the different touchpoints observed in chronological order.

Using the new format it was possible to calculate the distance between the sequences or CJs. The method used to do this was the optimal matching edit distance method, which was first introduced by Levenshtein [36]. With this method, the distance between two sequences is calculated by looking at the costs for substitution, insertion and deletion. The cost for insertion and deletion was set by using a predefined method, which makes these costs equal to the maximum value of the substitution cost matrix divided by two. Where the substitution cost matrix contains all the costs for substituting one state for another state of the alphabet. In order to calculate these substitution costs the "TRATE" method was used, which determines the cost using the probabilities that states follow each other. The costs are determined from the estimated transition rates as: $2 - p(s_i|s_j) - p(s_j|s_i)$, where $p(s_i|s_j)$ is the probability of observing state s_i at time t+1 given that state s_i has been observed at time t and $p(s_i|s_i)$ the other way around. The idea is to set a high cost when changes between s_i and s_i are rarely observed and lower cost when they are common [18]. In Table 2 an example is given based on two different CJs. Assuming that there is at least one transition of states that never occurs, the maximum value in the substitution matrix equals two. In that case, the insertion and deletion cost will equal one, since this value is equal to the maximum of the substitution matrix divided by two. In addition, we assume that the substitution cost for $4 \iff 5$, $7 \iff 10$ and $10 \iff 16$ equal 1.7, 1.75 and 1.9 respectively. In that case the cost for the substitutions equals 5.35 and the cost for the insertions is equal to 1, which brings the total cost or distance between these two journeys to 6.35.

Table 2. Calculating distance using touchpoints.

Seq1 18	5	47	10	16	16
Seq2 \perp	4	$4\ 7$	7	16	10
Cost 1	1.7	0 0	1.75	0	1.9

Within the context of CJs, there are large differences in trace length. With the purpose of dealing with these length differences we already proposed windows, but normalisation could also help in this situation. For the optimal matching distance, Abbott's normalization was used, which consists of dividing the distance by the length of the longest of the two sequences. This way the edit distance was corrected for the difference in journey length [18]. The result of the previous steps is a matrix holding the edit distance between each pair of CJ sequences.

With the resulting distance matrix of the previous steps it was possible to start clustering. Although here we want to propose an extension to alter the distances used to cluster. The main idea is not only to use the distance based on the touchpoint but also to use distance based on other information levels. In the data there is also the level indicating who initiated the touchpoint, which was added as initiator. Also the other manually added column, message, can serve as another level of information. For both levels the same steps were taken as for calculating the distance based on the touchpoints. This resulted in the presence of three different distances matrices, i.e., touchpoint, initiator and message. Using these three matrices, different scenarios were tested in order to form high-quality clusters.

In the proposed approach we bring together the matrices into one distance matrix before clustering. This can be done by using a weighted average of these matrices to form a total distance matrix. The reasoning behind this approach is that in this way the distances are altered to exaggerate the (dis)similarity. For example, take two separate CJs each from a different customer. There will be three different traces where the states are: touchpoint, message and the initiator. The base is the distance based on the touchpoints. When comparing the sequences of these customers there will be a certain distance based on these touchpoints in the trace. This distance was already calculated in Table 2 and equals 6.35. In Table 3 the same traces are displayed but this time using the message level. For the example we will assume that the substitution cost of Tour Op. \iff Generic equals 1.9. This is of course an assumption but in practice this cost can never be greater than 2 since the cost is calculated using the "TRATE"

method. When comparing the message level and the touchpoint level it could be that both customers were using different touchpoints but were looking for the same message. In the example this is the case with touchpoints 4 and 5 which are both information messages. Because of this similarity in the message the distance or cost between the two CJs decreases to 2.9 as shown in Table 3. Since there are less states at the message level compared to the touchpoint level the distance is likely to decrease, but when there are more state matches this drop will be more significant. The same goes for the traces looking at the initiator, an example is given in Table 4. Since at the level of initiator there are only two states, we assume that the maximum of the substitution matrix is 0.3, which means the costs for substitution and deletions equals 0.3 divided by 2. When these steps are performed for all observations they result in three different distance matrices. The next step is to bring them together into one matrix using a weighted average. Bringing together the different distances will always result in a distance smaller than the distance solely based on touchpoints, but for the journeys showing more similarities at the higher information levels the decrease will be bigger. By using this multi-level aggregation the goal is to correct the distance for the observed behaviour. The applied weights cannot be interpreted as being the importance, due to the different scales for the distances on the 3 levels as a result of the different number of states. These weights are solely a correction to search for the best quality clusters using these different information levels. When looking at the correlation of the calculated distances, the Spearman's rank correlation was used, since the data is not normally distributed. This resulted in a correlation of 0.905 between the touchpoint distance and the initiator distance, a 0.996 correlation between touchpoint distance and message distance and a correlation of 0.910 between initiator distance and message distance. Consequently, we can conclude that the correlation between the three distances is strong. This also means that most of the information is already captured in the touchpoint distance.

Table 3. Calculating distance using message.

Seq1	Ind. M.	Info	Info	Tour O	p. Tour Op.	Generic	Generic
Seq2	\perp	Info	${\rm Info}$	Tour O	p. Tour Op.	Generic	Tour Op.
Cost	1	0	0	0	0	0	1.9

Table 4. Calculating distance using initiator.

Seq1	Comp.	Cust.	Cust.	Cust.	Cust.	Cust.	Cust.
Seq2	\perp	${\rm Cust.}$	Cust.	Cust.	Cust.	${\rm Cust.}$	${\rm Cust.}$
Cost	0.15	0	0	0	0	0	0

5.2.3 Clustering the data

The resulting total distance matrix created during the previous steps can now be used to cluster the CJs. The aim is to bring together sequences, which show similar behaviour (small distance), to form homogeneous groups. The sequences in the other clusters should be as far away as possible from this sequence. The clustering methods tested are a agglomerative hierarchical clustering and a combination of the previous one followed by a k-medoids clustering algorithm, i.e., hybrid clustering. In this combination the first step is to perform a hierarchical clustering algorithm. Second, using the agglomeration schedule, multiple cluster quality measures can be calculated. Based on these measures the optimal number of clusters is identified. The third step is to deploy a k-medoids algorithm to optimize the results. The medoids are based on the results of the hierarchical clustering and the number of clusters identified. After the k-medoids algorithm is applied, the final clusters are reached.

To identify the quality of the clusters and the optimal number of clusters, four different quality measures are calculated. The Point Biserial Correlation (PBC) and the Hubert's Gamma (HG) are both coefficients measuring the capacity to reproduce the distance matrix. The PBC coefficient is based on the exact value, whereas the HG coefficient looks at concordance. This implies that, according to the HG coefficient, a cluster is valid if the distances between the groups are greater than those within the groups [18]. Both the PBC and HG take a value between -1 and 1 and a higher value means better clusters. The Hubert's C (HC) coefficient measures the gap between clusters obtained and the best cluster theoretically possible with the number of clusters and the distances [18]. The HC coefficient takes a value between 0 and 1, where you want to minimize the value. Lastly the average silhouette width (ASW) measures the coherence of assignments. A high coherence indicates high between-group distance and strong within-group homogeneity [18]. This measure takes values between -1 and 1, were intervals are identified with their own meaning which are reproduced in Table 5. The four measures are plotted on a graph, in which, the options from 2 to 10 clusters are projected with their corresponding values for the metrics (see Figure 2). The maximum number of clusters created is set to 10, since one should always keep in mind the practical feasibility. The optimal number of clusters can then be visually derived from the graph.

Table 5. Orders of magnitude for interpreting the ASW measure.

ASW	Interpretation
0.71 - 1.00	Strong structure identified.
0.51 - 0.70	Reasonable structure identified.
0.26 - 0.5	Structure is weak and could be artificial.
<= 0.25	No structure.

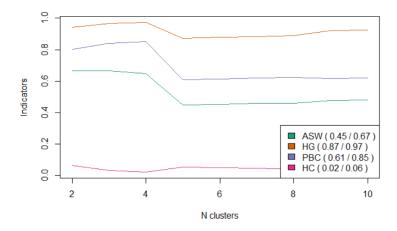


Fig. 2. Example quality metrics graph.

5.3 Analyzing clusters

The dataset is now enriched with the corresponding cluster for each customer. Using the old format where the dataset consisted of the UserID, the timestamp and the touchpoint, an eventlog is created and expanded with the cluster variable. Creating the eventlog allows for process mining/discovery techniques to be applied to further study the behaviour of the different clusters identified. This is done by looking at structural log metrics, studying touchpoint occurrence and generating directly follows graphs.

6 Results

6.1 Scenarios

For all tested scenarios, the first 100 000 observations of the dataset were used. Also, the CJ traces were truncated at 230 observations. This 230 was determined by looking at the entire dataset, where the shortest 75% of the traces from the dataset had a length between 0 and 230. Both choices had to be made for performance reasons. The exact specifications of some tested scenarios can be found in Table 6. The scenarios shown in this table are the most important ones from the iterative process towards the most qualitative clusters.

For the first scenario the weights that form the total distance matrix were respectively 0.5 for the touchpoints and 0.25 for initiator and message. The ASW of 0.148, relatively small PBC and HG and large HC values indicates the absence of structure, with no real cluster formation. As a result of trying to optimize the weights to form the distance matrix, scenario 2 was created. Since the ASW equals 0.406 and improvements were made for the PBC, HG and HC values this

Table 6. Scenario results.

G	Weight	Weight	Weight	Normali-	XX7° 1	Hybrid	Number of	PBC	HG	НС	ASW
Scen.	touchpoint					clustering	clusters	[-1,1]	[-1,1]	[0,1]	[-1,1]
1	0.5	0.25	0.25	✓	Х	✓	8	0.508	0.665	0.175	0.148
2	0.05	0.85	0.1	✓	X	✓	8	0.773	0.970	0.026	0.406
3	0.05	0.85	0.1	\checkmark	\checkmark	\checkmark	6	0.731	0.929	0.069	0.366
4	0.05	0.85	0.1	X	X	X	4	0.761	0.961	0.014	0.530
5	0.05	0.85	0.1	X	\checkmark	X	5	0.765	0.971	0.011	0.589
6	0.5	0.3	0.2	X	\checkmark	×	5	0.809	0.942	0.054	0.597
7	0.5	0.4	0.1	X	\checkmark	X	4	0.851	0.970	0.021	0.648

indicated that there is some cluster formation, but this still could be artificial. With the goal to further improve the quality and to overcome clustering based on loyalty, windows were introduced. The results of scenario 3 use the same settings as those in scenario 2, with the exception of windows that are applied. Although the quality metrics are still acceptable, they are deteriorated compared to those of scenario 2. It again implies that there is some form of clustering, but could still be artificial.

Through further analysis, it appeared that by omitting normalisation, the quality of the clusters increased notably. Also when no longer using normalisation, the results proved to be stronger solely with hierarchical clustering instead of hybrid clustering. This new approach is summarised in scenario 4. The ASW of 0.53 for the first time indicates reasonable clustering. The PBC and HG values are very close to those of scenario 2, the HC value is almost halved, also indicating an improvement. By disregarding normalisation, the difference in length is no longer controlled for. Also, clustering according to loyalty can be a problem again since no windows are used in scenario 4. Using the quarter windows again, we can control for both shortcomings with the goal to improve the results.

Scenario 5 (i.e. scenario 4 with windows applied) shows an overall improvement of all the metrics. With the goal to improve the quality even further in this situation, the weights used for the creation of the total distance matrix were reviewed. This resulted in scenarios 6 and 7 where both are still a definite improvement compared to scenario 5, this with the exception of the HC metric which rises again. However, this metric is still relatively small, especially in scenario 7, and still indicates strong clusters. In the end, scenario 7 gives the most qualitative clusters, amounting to four.

6.2 Analyzing clustering results

6.2.1 Structuredness

The most qualitative clusters were created in scenario 7, therefore we will use them for the further analyses. On a more technical level the clusters are analysed using structural log metrics. These metrics allow to position the clusters and the systems producing them in a domain-independent fashion, since these measures are derived from the log itself and do not depend on semantics or external do-

main knowledge [22]. Remember that the first 100 000 observations were used to create the clusters. In Table 7 the second column displays the distribution of these touchpoints over the clusters. The support column reflects how many CJs the cluster contains. Based on these two columns one can already derive that the overall CJ length in cluster three will be higher compared to the other clusters since it only contains 96 CJs but has the highest number of touchpoints. It is also the case that cluster 1 represents as much as 80% of total support. In order to calculate the level of detail the summation of all the distinct activities per CJ, which is also called the variety of a CJ, was divided by the support of the cluster. This results in a mean number of distinct activities per CJ for a cluster. The level of detail therefore makes abstraction of repetitions within a CJ, which would count in the mean CJ length. While the level of detail decreases slightly for cluster 1, it increases considerably for the other three clusters. These clusters therefore contain more CJs that do not just repeat the same touchpoint, but where the customer uses several touchpoints. Lastly, the structure of an event log describes the amount of observed behaviour, as compared to the amount of theoretically possible behaviour [22]. The lower the structure the more likely every touchpoint is connected to every other touchpoint. For all the clusters the structure metric increases, which supports the conclusion that the identified clusters are from a certain quality.

Cluster Number of touchpoints Support Level of detail Structure All 100000 2.672 0.403 1579 1 26828 1302 2.253 0.548 2 131 214624.099 0.5683 37111 96 5.229 0.600 4 14599 50 4.940 0.564

Table 7. Cluster structuredness measures.

6.2.2 Touchpoint frequency and directly follows graph

The goal of the clustering is to deliver more valuable insight compared to the representative journey already existing. Using existing process mining techniques we visualise the CJ. First a directly follows graph (i.e. customer journey map) is created with BupaR [32], of all the data used to do the clustering, which is visualised in Figure 3. This directly follows graph is able to represent the 90% most common CJs in the analysed data. When looking at the touchpoints the absolute numbers are given on the graph, for the flow relations the relative percentage is displayed. The colours also indicate the importance, where darker colours stand for more frequent behaviour. From Figure 3 it follows that the touchpoints 7 and 1 are the most frequent touchpoints, they represent respectively 35% and 32,7% of all the touchpoints. All the other touchpoints have an occurrence of less than 10% in the dataset.

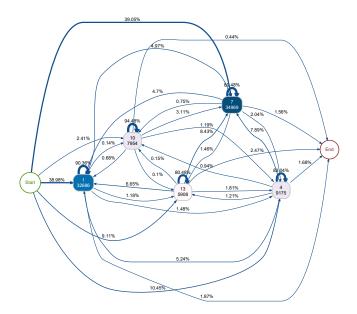


Fig. 3. 90% graph of all 100 000 touchpoints.

When looking at the clusters of scenario 7, four different clusters were identified. Concerning the number of touchpoints cluster 1 represents 26,8%, cluster 2 represents 21,5% and cluster 3 and 4 respectively 37,1% and 14,6% of all the touchpoints in the clustered data. For each cluster the same 90% directly follows graph was created, which are displayed in Figures 4 until 7. In Figures 8 until 11 a comparison was made for the ten most frequent touchpoints, which represent 98.5%, of the 100 000 used touchpoints. In these figures the baseline represents the frequency observed over all touchpoints together, the green bars indicate a relative increase in the presence of a touchpoint for that cluster and the red bars a decrease in the relative presence. In cluster 1, visible in Figure 4, the touchpoint 1 represents 39,6%, which is an increase with 6.9% and touchpoint 7 represents 31,7% which is a decrease with 3.3%. When comparing the general directly follows graph with the one of cluster 1, both touchpoint 1 and 7 remain the most important touchpoints. Figure 8 also indicates the decrease of touchpoint 4 and 10.

Looking at the important touchpoints, still touchpoint 1 and 7 represent respectively 27,5% and 23,4% of all touchpoints in cluster 2, which is a decrease compared to the level of all touchpoint. Figure 9 also shows an increase in the presence of touchpoints 2, 4 and 5 and a slight increase for the touchpoints 13 and 14. The reason for the third important touchpoint in Figure 5 is thus confirmed by the 8.8% increase, which is visible in Figure 9. For cluster 2 the 90% graph was more complex than the graph for all the data. This can be the result of having multiple touchpoints, beside touchpoint 1 and 7, with similar relative importance without one or a few of them really standing out.

Looking at cluster 3 the presence of touchpoint 7 notably increases, with almost 20%, to 53,9% and the presence of touchpoint 1 drops even further to only 13,7% of all touchpoints in this cluster. Within this cluster the distinctive touchpoint, is touchpoint 10, representing 13,7%, which is a 5% increase compared to the general level.

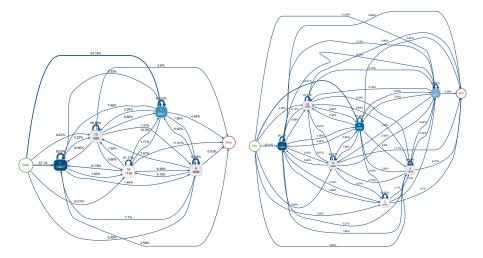


Fig. 4. 90% graph cluster 1.

Fig. 5. 90% graph cluster 2.

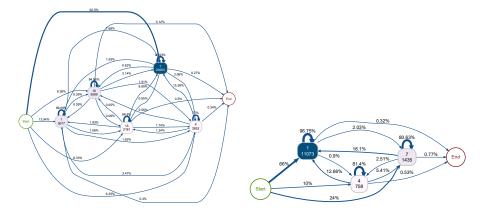


Fig. 6. 90% graph cluster 3.

Fig. 7. 90% graph cluster 4.

Finally analyzing cluster 4, the map considerably decrease in complexity compared to the general map. For this cluster there is only one touchpoint with an occurrence above 10%, which is touchpoint 1 that occupies as much

as 75% of all touchpoints in the entire cluster. This is reflected in Figure 11 by showing an increase above 40%. Touchpoint 7 shows a strong decrease, but also for touchpoints 4, 10 and 13 this decrease is clearly visible.

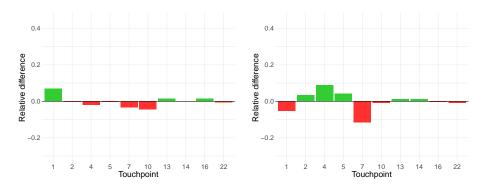


Fig. 8. Touchpoint difference cluster 1.

Fig. 9. Touchpoint difference cluster 2.

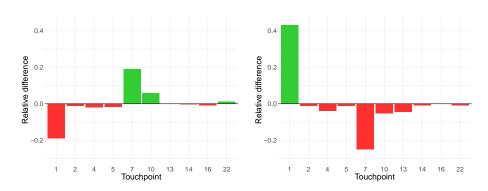


Fig. 10. Touchpoint difference cluster 3. Fig. 11. Touchpoint difference cluster 4.

Further interpreting the differences in the important touchpoints among clusters, one could make the following statements:

- Cluster 1 is the cluster which most closely resembles the general picture of the selected data, where the customers both take a look at the website for accommodations and also looks at websites of competitors.
- The same goes for cluster 2, where they also look at websites for accommodations and the website of competitors. Although the usages of both touchpoints drops compared to cluster 1, but they use an other alternative by looking at the information and comparison websites as well.

- Cluster 3 is highly reliant on the tour operator which they both use to look at competitors in the first place, but also for a small portion to look at the brand the company of this study represents. They also use the website to look at accommodations.
- The cluster which contains the most loyal clients is cluster 4. Given that in this cluster the touchpoint most heavily used is the website to search for accommodations and other touchpoints only occur sporadic, were no touchpoints lead them to competitors.

7 Discussion

By using the data that contains the observed behaviour of customers it is possible to differentiate between customer profiles. Like Terragni and Hassani [49] suggested there are different approaches to clustering CJs. The two methods translating the data into a vector-space are more feature orientated, with the drawback of loosing the dynamics. Therefore, a method that still captures the dynamics of the behaviour and thus holds the CJ together as a sequence was preferred. By using the optimal matching edit distance it was possible to create different clusters successfully. Proving this concept plays an important role in showing the potential of the CJA field, not only for scientific research but also to use in practice. When further developed, the approach could deliver another alternative for marketeers to create different customer profiles. Nowadays this is often done by using demographic data or purchase data. Often data concerning the touchpoints is already collected, this in the context of CRM and is classified as relational data [54]. Understanding the different clusters can be the basis towards selecting the best marketing tactics, to classify new customers into the existing cluster or even to start predicting the next touchpoint of customers.

Although the approach presented in this paper shows its potential, there are still some limitations. When using the optimal matching distance, determining the costs for insertion and deletions as well as the cost for substitution are very difficult. Therefore, in this paper, standard methods were used which take the probabilities of the touchpoints into account. Although these methods should deliver good results, it could well be that the quality of the clusters can be further improved by testing different costs. As a result of using different costs it could be beneficial to again consider the weights used to create the total distance matrix. In this case the applied weights deliver the best results, but this could not be the case with other data or with the other parameters tweaked differently. In other words, we recommend to further research the costs and settings to improve the quality of the clusters. The data used in this study does not perfectly match the proposed definition. Nevertheless, the clustering approach is easily extendable to a better match with the definition. Explicitly considering the channel as a third level to calculate the distance could be further research for better clustering. With the focus towards the application of the approach in businesses, analysis and visualisation should be more tailored towards marketeers. Therefore, studying the field of customer segmentation can indicate the needs of those marketeers and the way they try to use these insights. Then transferring these needs to the domain of CJA and tailor the outcome towards the decision-making process of the end users. Also the concept of stages, introduced in the CJ definition, can get applied in the future to make the outcome more comprehensible.

8 Conclusion

Overall the contribution of our work is twofold: we introduced a formal definition of the CJ to use in the field of CJA and proved the potential of clustering CJs to make them more valuable for decision-making within marketing.

In absence of a clear definition of the CJ in literature, we consolidated different opinions and perspectives into one formal definition. To build this definition, a touchpoint with following components: type, message and channel was defined. When creating a sequence of touchpoints and linking the sequence with a customer, a CJ is created. We also defined an optional extension to the definition by using a mapping function for the touchpoints which divides the CJ into stages. Secondly we tried to extend the value of the actual CJ by using clustering. Based on the traces the edit distance was calculated. A new approach using multiple information levels (e.g. touchpoint, initiator and message) to calculate that distance was proposed. Using that resulting distance matrix of the optimal matching edit distance, an hierarchical approach as well as an hybrid approach was tested to clusters. In order to deal with the substantial length differences between CJs, normalisation and windows were considered. The best quality clusters were found when using only hierarchical clustering with no normalisation of the distance but using windows, in this case based on the quarters. The quality was identified by using the Point Biserial correlation, Hubert's Gamma, Hubert's C and the average silhouette width. In the analysis the use of structural log metrics confirmed that the clusters found were of a certain quality. Using directly follows graphs and comparing touchpoint occurrence revealed differences in the importance of certain touchpoints among the clusters. These differences allowed to make some high level conclusions about the customers in a certain cluster.

Acknowledgements

Throughout the writing of this paper I have received a great deal of support and assistance.

I would first like to thank my supervisor, Dr. Gert Janssenswillen, whose expertise was invaluable in conducting this research. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to acknowledge Prof. Dr. Benoit Depaire, for his valuable guidance throughout the process of defining the customer journey. You provided me with the tools that I needed to choose the right direction and successfully complete this chapter.

I would also like to thank my cousin Dr. Jente Broeckx, for his time and effort to evaluate this paper from an external perspective. I also want to thank him to inspire me throughout my youth to pursue a university degree.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my girlfriend, Zoë Grossi, who provided happy distractions to rest my mind outside of my research.

References

- E. Anderl, I. Becker, F. Von Wangenheim, and J. H. Schumann. Mapping the customer journey: Lessons learned from graph-based online attribution modeling. *International Journal of Research in Marketing*, 33(3):457–474, 2016.
- N. Barwitz and P. Maas. Understanding the omnichannel customer journey: determinants of interaction choice. *Journal of interactive marketing*, 43:116–133, 2018.
- G. Bernard. Process Mining-based Customer Journey Analytics. PhD thesis, Université de Lausanne, 2020.
- 4. G. Bernard and P. Andritsos. A process mining based model for customer journey mapping. In Forum and Doctoral Consortium Papers Presented at the 29th International Conference on Advanced Information Systems Engineering (CAiSE 2017), pages 49–56. CEUR Workshop Proceedings, 2017.
- G. Bernard and P. Andritsos. Cjm-ab: Abstracting customer journey maps using process mining. In *International Conference on Advanced Information Systems* Engineering, pages 49–56. Springer, 2018.
- G. Bernard and P. Andritsos. Contextual and behavioral customer journey discovery using a genetic approach. In *European Conference on Advances in Databases and Information Systems*, pages 251–266. Springer, 2019.
- 7. G. Bernard and P. Andritsos. Discovering customer journeys from evidence: a genetic approach inspired by process mining. In *International Conference on Advanced Information Systems Engineering*, pages 36–47. Springer, 2019.
- 8. A. Bifet, B. Pfahringer, J. Read, and G. Holmes. Efficient data stream classification via probabilistic adaptive windows. In *Proceedings of the 28th annual ACM symposium on applied computing*, pages 801–806, 2013.
- 9. A. Bookstein, V. A. Kulyukin, and T. Raita. Generalized hamming distance. *Information Retrieval*, 5(4):353–375, 2002.
- 10. R. J. C. Bose and W. M. Van der Aalst. Context aware trace clustering: Towards improving process mining results. In *proceedings of the 2009 SIAM International Conference on Data Mining*, pages 401–412. SIAM, 2009.
- 11. D. Buscaldi, R. Tournier, N. Aussenac-Gilles, and J. Mothe. Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 552–556, 2012.
- 12. C. Chiu. A case-based customer classification approach for direct marketing. Expert Systems with Applications, 22(2):163–168, 2002.
- 13. S. Clatworthy. Service innovation through touch-points: the at-one touch-point cards. In *Proceedings of 2nd Service Design and Service Innovation Conference*, pages 25–38, 2010.
- 14. S. Clatworthy. Service Innovation Through Touch-points: Development of an Innovation Toolkit for the First Stages of New Service Development. *International Journal of Design*, 5(2):15–28, 2011.
- 15. M. H. H. Cordewener. Customer journey identification through temporal patterns and Markov clustering. Master's thesis, Eindhoven University of Technology, 2016.
- M. D'Arco, L. Lo Presti, V. Marino, and R. Resciniti. Embracing AI and Big Data in customer journey mapping: from literature review to a theoretical framework. *Innovative Marketing*, 15(4):102–115, 2019.

- A. Følstad and K. Kvale. Customer journeys: a systematic literature review. Journal of Service Theory and Practice, 28(2):196–227, 2018.
- 18. A. Gabadinho, G. Ritschard, N. S. Müller, and M. Studer. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37, 2011.
- A. Gabadinho, G. Ritschard, M. Studer, and N. S. Mueller. Summarizing sets of categorical sequences: selecting and visualizing representative sequences. In *International Conference on Knowledge Discovery and Information Retrieval*, pages 94–106, 2009.
- 20. M. George and K. L. Wakefield. Modeling the consumer journey for membership services. *Journal of Services Marketing*, 32(2):113–125, 2018.
- 21. S. Goyat. The basis of market segmentation: a critical review of literature. European Journal of Business and Management, 3(9):45–54, 2011.
- C. W. Günther. Process mining in flexible environments. PhD thesis, Eindhoven University of Technology, 2009.
- 23. T. H. Davenport. How strategists use "big data" to support internal business decisions, discovery and production. *Strategy & Leadership*, 42(4):45–50, 2014.
- R. Halvorsrud, K. Kvale, and A. Følstad. Improving service quality through customer journey analysis. *Journal of service theory and practice*, 26(6):840–867, 2016.
- M. Harbich, G. Bernard, P. Berkes, B. Garbinato, and P. Andritsos. Discovering customer journey maps using a mixture of markov models. In SIMPDA, pages 3–7, 2017.
- 26. I. M. Haugstveit, R. Halvorsrud, and A. Karahasanovic. Supporting redesign of c2c services through customer journey mapping. In Service Design Geographies. Proceedings of the ServDes. 2016 Conference, pages 215–227. Linköping University Electronic Press, 2016.
- D. Herhausen, K. Kleinlercher, P. C. Verhoef, O. Emrich, and T. Rudolph. Loyalty formation for different customer journey segments. *Journal of Retailing*, 95(3):9– 29, 2019.
- 28. M. Heuchert, B. Barann, A.-K. Cordes, and J. Becker. An is perspective on omni-channel management along the customer journey: Development of an entity-relationship-model and a linkage concept. In *Multikonferenz Wirtschaftsinformatik*, pages 435–446, 2018.
- 29. A. Hiziroglu. Soft computing applications in customer segmentation: State-of-art review and critique. Expert Systems with Applications, 40(16):6491–6507, 2013.
- 30. S. Hogan, E. Almquist, and S. E. Glynn. Brand-building: finding the touchpoints that count. *Journal of Business Strategy*, 26(2):11–18, 2005.
- 31. C. Ingo Berendes, C. Bartelheimer, J. Hendrik Betzing, and D. Beverungen. Data-driven customer journey mapping in local high streets: A domain-specific modeling language. In *Proceedings of the 39th International Conference on Information Systems*, pages 1–8, 2018.
- G. Janssenswillen, B. Depaire, M. Swennen, M. Jans, and K. Vanhoof. bupaR: Enabling reproducible business process analysis. *Knowledge-Based Systems*, 163:927– 930, 2019.
- 33. M.-S. Kim, K.-Y. Whang, J.-G. Lee, and M.-J. Lee. n-gram/2l: A space and time efficient two-level n-gram inverted index structure. In *Proceedings of the 31st international conference on Very large data bases*, pages 325–336, 2005.
- 34. C. Lazaris and A. Vrechopoulos. From multi-channel to "omnichannel" retailing: review of the literature and calls for research. In 2nd International Conference on Contemporary Marketing Issues, (ICCMI), pages 1–6, 2014.

- 35. K. N. Lemon and P. C. Verhoef. Understanding customer experience throughout the customer journey. *Journal of marketing*, 80(6):69–96, 2016.
- 36. V. Levenshtein. Levenshtein distance, 1965.
- 37. J. Lies. Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(5):134, 2019.
- 38. G. Liu, T. T. Nguyen, G. Zhao, W. Zha, J. Yang, J. Cao, M. Wu, P. Zhao, and W. Chen. Repeat Buyer Prediction for E-Commerce. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 155–164, 2016.
- T. Ludwig, X. Wang, C. Kotthaus, S. Harhues, and V. Pipek. User narratives in experience design for a b2b customer journey mapping. *Mensch und Computer* 2017-Tagungsband, pages 193–201, 2017.
- A. Mosquera, E. J. Ayensa, C. O. Pascual, and Y. S. Murillo. Omnichannel Shopper Segmentation in the Fashion Industry. *Journal of Promotion Management*, 25(5):681–699, 2019.
- A. Nairn and P. Berthon. Creating the Customer: The Influence of Advertising on Consumer Market Segments – Evidence and Ethics. *Journal of Business Ethics*, 42(1):83–100, 2003.
- 42. S. Nakano and F. N. Kondo. Customer segmentation with purchase channels and media touchpoints using single source panel data. *Journal of Retailing and Consumer Services*, 41:142–152, 2018.
- S. A. Neslin, D. Grewal, R. Leghorn, V. Shankar, M. L. Teerling, J. S. Thomas, and P. C. Verhoef. Challenges and opportunities in multichannel customer management. *Journal of service research*, 9(2):95–112, 2006.
- 44. H.-L. Nguyen, Y.-K. Woon, and W.-K. Ng. A survey on data stream clustering and classification. *Knowledge and Information Systems*, 45(3):535–569, 2015.
- 45. A. Richardson. Using Customer Journey Maps to Improve Customer Experience. *Harvard Business Review*, 15(1):2–5, 2010.
- M. S. Rosenbaum, M. L. Otalora, and G. C. Ramírez. How to create a realistic customer journey map. *Business Horizons*, 60(1):143–150, 2017.
- A. Siebert, A. Gopaldas, A. Lindridge, and C. Simões. Customer Experience Journeys: Loyalty Loops Versus Involvement Spirals. *Journal of Marketing*, 84(4):45–66, 2020.
- 48. M. Song, C. W. Günther, and W. M. P. van der Aalst. Trace Clustering in Process Mining. In *Business Process Management Workshops*, pages 109–120, 2009.
- 49. A. Terragni and M. Hassani. Analyzing customer journey with process mining: From discovery to recommendations. In 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pages 224–229, 2018.
- A. Terragni and M. Hassani. Optimizing customer journey using process mining and sequence-aware recommendation. In *Proceedings of the 34th ACM/SIGAPP* Symposium on Applied Computing, pages 57–65, 2019.
- J. L. Viegas, S. M. Vieira, R. Melício, V. M. F. Mendes, and J. M. C. Sousa. Classification of new electricity customers based on surveys and smart metering data. *Energy*, 107:804–817, 2016.
- Weinan Wang and O. R. Zaiane. Clustering Web sessions by sequence alignment. In Proceedings. 13th International Workshop on Database and Expert Systems Applications, pages 394–398, 2002.
- 53. R.-S. Wu and P.-H. Chou. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3):331–341, 2011.

- 54. D. Zahay, J. Peltier, D. E. Schultz, and A. Griffin. The Role of Transactional versus Relational Data in IMC Programs: Bringing Customer Data Together. *Journal of Advertising Research*, 44(1):3–18, 2004.
- 55. L. G. Zomerdijk and C. A. Voss. Service design for experience-centric services. *Journal of service research*, 13(1):67–82, 2010.
- 56. Q. Zu, T. Wu, and H. Wang. A Multi-factor Customer Classification Evaluation Model. *COMPUTING AND INFORMATICS*, 29(4):509–520–520, 2012.