Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

School of Engineering and Sciences

Characterisation of visitors and description of their navigation behaviour using Web Log Mining techniques

A thesis presented by

Alicia Huidobro Espejel

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Atizapán de Zaragoza, Estado de México, Feb, 2021

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

School of Engineering and Sciences

The committee members, hereby, certify that have read the thesis presented by Alicia Huidobro Espejel and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Science.

Dr. Raúl Monroy Borja
Tecnológico de Monterrey, Campus Estado de México
Principal Advisor

Dra. Bárbara Cervantes González
Tecnológico de Monterrey, Campus Estado de México
Co-Advisor

Dr. Octavio Loyola-González
Tecnológico de Monterrey, Campus Puebla
Committee Member

Dr. Mario Graff Guerrero
Centro de Investigación e Innovación en Tecnologías
de la Información y Comunicación (INFOTEC)
Committee Member

Dr. Raúl Monroy Borja
Director de Programa de Graduados en Computación, región CDMX
School of Engineering and Sciences

Atizapán de Zaragoza, Estado de México, Feb, 2021

Declaration of Authorship

I, Alicia Huidobro Espejel, declare that this thesis titled, ”Characterisation of visitors and description of their navigation behaviour using Web Log Mining techniques” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Alicia Huidobro Espejel
Atizapán de Zaragoza, Estado de México, Feb, 2021

Dedication

I dedicate my dissertation work to my beloved Rodri and Axel for being the joy and light of my life. To Adán for being my rock. To my parents and siblings for being the best example of determination and self-improvement.

Acknowledgements

My deepest gratitude to my advisors, Dr. Raúl Monroy Borja and Dra. Bárbara Cervantes González, who were more than generous with their expertise and precious time. Their suggestions and guidance were essential for the successful development of this research.

Special thanks to Dr. Raúl Monroy Borja for his mentoring. I sincerely appreciate the opportunity to dive me into the fascinating world of Data Science and Machine Learning.

I also thank the Committee Members, Dr. Octavio Loyola-González and Dr. Mario Graff Guerrero, for their valuable comments and suggestions to improve this work.

Thanks to Tecnológico de Monterrey and CONACYT for supporting this research.

Last but not least, thanks to NIC México for sharing data that was fundamental for this work. This data does not jeopardise the confidentiality of its clients in any way. Thanks to the Marketing and Information Technology teams at NIC México, for providing feedback during this project.

Characterisation of visitors and description of their navigation behaviour using Web Log Mining techniques

by

Alicia Huidobro Espejel

Abstract

The value of a company's website depends on visitors performing actions that add value for the company. Those actions are called conversions. We present techniques for both characterising website visitors in terms of the conversions they make, and describing their navigation behaviour in an abstract way, with the aim of making them more amenable to interpretation. Existing web analytics techniques have not been designed to highlight the distinguishing characteristics of a class of visitors. There are no approaches for characterising classes of visitors that take into account specific business goals; further, the navigation behaviour of a visitor, let alone a class of visitors, is conveyed as a sequence of visited pages, without giving this an abstract meaning. In this thesis, we introduce a means of characterising website visitors. To find what the different segments of visitors have or do not have in common, we first separate visitor sessions in terms of conversions and then for each class we mine patterns to contrast one another. We also introduce a simplified description of visitor navigation behaviour. Our technique works by identifying subsequences of visited pages of common occurrence, called "rules", and then by shrinking a session replacing those rules with a symbol that is given a representative name. Further, we extended this to an entire class of visitors, creating a graph that collects the class sessions, summarising the class navigation behaviour and enabling an easier contrast of classes. Our results show that a few patterns are enough to characterise a visitor class; since each class is associated with a conversion, an expert can easily draw conclusions as to what makes two classes different from one another. Also, with our abstract representation, a session can be shrinked so that the behaviour of an entire visitor class can be depicted in a moderately small graph. Further work is concerned with incorporating information from other sales channels and completing the analysis provided by existing techniques.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Companies are constantly looking for their website to stand out [18]. However, website visits do not necessarily represent a profit. To increase website profitability, it is necessary to identify which actions of the visitors represent a business benefit [24, 89, 87, 96]. Those actions are called conversions. Examples of conversions are rating a product, making a purchase, filling in a contact form, making a donation, etc. Marketing experts design strategies to increase the occurrence of conversions. To improve the effectiveness of those strategies, it is valuable to know the characteristics of the visitors who perform a given conversion and understand their navigation behaviour. Obtaining this knowledge requires the analysis of the website data.

To analyse website data, there are different Web Analytics Solutions (WAS) [53, 108, 2, 5, 6, 11, 14, 12, 15]. They include commercial software, like Google Analytics, and different approaches found in the literature. Existing WAS provide useful information about the characteristics of visitors and how they navigate the website. Nevertheless, WAS have three limitations for characterising visitors and describing their navigation behaviour:

1. The characterisation of visitors consists of finding the properties of members of a class that cannot be associated with members of other classes [48]. Commercial software does not provide a characterisation of visitors. Instead, it allows the analysis of characteristics separately. These are shown in spread tables and charts [2, 5]. This hinders to find out the set of distinguishing characteristics for a given class of visitors. Approaches found in the

literature have centred around finding clusters of visitors using machine learning algorithms [98, 47, 54, 58]. Nevertheless, they do not provide a characterisation for classes of visitors.

2. Regarding the navigation behaviour of visitors, commercial software provides a page-by-page report [2, 5]. This level of detail produces huge graphs that are difficult to analyse and compare. Some approaches in the literature analyse the sequence of visited pages [104, 69, 94, 42, 41]. However, they do not provide a high-level description of the navigation behaviour. They are limited to cluster visitors based on different criteria, for example, the longest common subsequence of visited pages [19].

3. To characterise visitors and describe their navigation behaviour, we propose to group visitors according to the conversions that they perform. However, commercial software hinders the identification of classes of visitors based on business goals [10, 13]. Instead, it allows analysing some common market segments (for example, male and female visitors) [2, 5]. In the literature, Machine Learning is widely used to analyse website traffic. Especially to find clusters of visitors [92, 98, 47, 54, 58]. However, conversions have not been previously used to classify visitors.

Based on the WAS limitations, we established two main objectives for this research. The first objective is to characterise website visitors according to the conversions they make. The second objective is to describe the navigation behaviour of a whole class of visitors. To achieve these objectives, we used web log files as input data, and we explored alternative approaches that allowed us to overcome the limitations of WAS. Our methodology is summarised below.

Characterisation of visitors according to conversions

To characterise visitors, we need to find the distinguishing characteristics for different classes of visitors [48, 52]. We started by grouping the activities performed by each unique visitor in a period. This group of activities is commonly known as "session" [2, 5, 86, 85, 33, 64, 44]. Web traffic contains sessions performed by humans and sessions performed by bots [2, 5, 99, 21]. We eliminated bots sessions because we are interested in characterising human visitors. We grouped the human sessions according to the conversions they performed. Each group of sessions

corresponds to a class. For each class, we obtained a set of contrast patterns. A contrast pattern describes the properties of members of a class that cannot be associated with members of other classes [73, 29]. Therefore, the contrast patterns obtained for each class of visitors characterise the visitors from that class.

Description of the navigation behaviour of visitors

To describe the navigation behaviour of a whole class of visitors, we started by representing sessions as the sequence of visited pages. We extracted the sequences of pages that are most frequent among the visitors from a given class. We called those sequences rules. Then, we named those rules and used them to create a reduced representation of each session. A forty-page session could be reduced, for example, to two rules. Finally, we created a graph with the reduced sessions of the class. With this method, we reduce hundreds of nodes and edges into a simplified graph that captures the navigation behaviour of a whole class of visitors. It also allows to contrast classes.

Through this Chapter, we describe the previous concepts in more detail. In Section 1.1, we explain marketing concepts that are fundamental throughout this document. In section 1.2, we describe the problem that motivated our work. In Section 1.3 we present our hypothesis. Then, in Section 1.4, we explain the objectives of the research. Lastly, in Section 1.5, we present the methodology to achieve those goals.

1.1 Preliminary concepts

In this section, we explain two marketing concepts that are at the core of our proposal: “conversion” and “sales funnel”. We also describe how these terms relate to each other.

Conversion

Marketing experts can track numerous website metrics. However, in the end, the website value depends on its contribution to business objectives. Thus, marketing experts need to identify specific visitor actions that contribute to business objectives. In the marketing field, those actions are called conversions [82, 36, 83]. Examples of website conversions are: to pay for a product or a

service, to fill in a form with contact details, or to post a positive product review. Conversions are part of a broader process that we describe next.

Sales funnel

Few people make a payment the first time they visit a website. Usually, customers follow a journey. From the discovery of a brand to the frequent purchase of a product [24, 89, 87, 96]. The steps of that journey are known, in the marketing field, as the "sales funnel." The most traditional sales funnel model is four-step: Attention, Interest, Desire, and Action [91, 93, 65]. It is known as AIDA model. However, each company can design its sales funnel with different steps, based on its business objectives, product characteristics, and industry rules.

The relevance of conversions in the sales funnel

Marketing experts design strategies to move customers through the sales funnel steps. For example, to get new visitors interested in the product (move them from attention to interest), marketing experts add valuable content on the website (e.g. blog posts, video tutorials, or technical reports). Marketing strategies are usually different for each sales funnel step [91, 93, 65]. The effectiveness of these strategies is usually measured through conversions [77, 84, 8]. For example, to determine the effectiveness of technical reports, we could measure the number of website visitors that downloaded them. Several conversions can refer to the same sales funnel step. Marketing experts need to measure the effectiveness of both each sales funnel step and each conversion.

Having clarified the previous marketing concepts, we can explain the problem that motivated this research.

1.2 Problem statement

The more knowledge a company has about visitors, the more effective its marketing strategies will be [100, 88, 82]. Therefore, it is valuable to know the characteristics of the visitors who perform desired actions on a website. It is also valuable to understand how they navigate the website [102, 83, 20]. This knowledge has to be obtained from the huge amount of data that is stored

on a website [20, 67, 25]. Web Analytics Solutions (WAS) are widely used and provide useful metrics [53, 108, 2, 5, 6, 11, 14, 12, 15]. However, they have some limitations for characterising visitors in terms of the conversions they make and describing their navigation behaviour. Below we describe three limitations that motivated our work:

1. Characterisation of visitors

To characterise visitors we need to find out the relevant characteristics of visitors who performed a given conversion [48, 52]. Nevertheless, commercial software shows the characteristics of visitors in scatter tables and charts [2, 5]. It allows to analyse characteristics independently, but it does not provide a characterisation as such. For example, if 30% of the visitors are from México, and 25% of visitors visit on Saturday. It is not easy to see if a relevant percentage of visitors has both characteristics. Approaches found in the literature do not characterise different classes of visitors either. They have centred around clustering visitors based on different criteria, for example, the frequently visited pages [98, 47, 54, 58].

2. Description of the navigation behaviour of visitors

We are interested in extracting the representative navigation behaviour of a whole class of visitors. Nevertheless, web sites could have hundreds of pages and thousands of visitors [106, 60]. It produces a huge number of different navigation paths [60]. Therefore, it is a challenge to provide a high-level description of the representative navigation behaviour of visitors. Commercial software shows different metrics about single pages (for example, the characteristics of the visitors who visited a given page) [2, 5]. It also provides the page-by-page sequence followed by visitors. Nevertheless, this level of detail produces huge sequences of pages that are hard to analyse. The navigation behaviour of a whole class of visitors may result in hundreds of pages to visualise. The comparison of different classes of visitors would be even more difficult. Approaches found in the literature analyse the navigation behaviour to find clusters of visitors but do not provide a high-level description of the navigation behaviour [78, 52, 51, 44, 94, 19].

3. Analysis of different classes of visitors

Regarding the classification of visitors, we are interested in grouping visitors according to the conversions they make. Commercial software allows comparing the characteristics of common market segments [2, 5]. For example, male and female visitors. Nevertheless, this hinders the analysis of additional classes of visitors [10, 13]. In literature, the use of different classes of visitors is common. However, we found no approaches in which classes of visitors are based on conversions.

1.3 Hypothesis

In this research, we validate a twofold hypothesis. First, by using classes of visitors based on conversions we can obtain a set of distinguishing patterns that characterise each class of visitors in a way that is closer to the language of marketing experts. An example of the expected characterisation could be that most visitors who make a payment are from México visit on Thursday and request online help. Second, we expect that by identifying sub-sequences of visited pages that appear frequently in visitor sessions of the same class, we can transform a session into a shorter sequence in which each frequent sub-sequence has been replaced by a symbol representing a general activity (which gets rid of a lot of the burden an expert needs to consider to understand visitor behaviour). For example, a twenty-page sequence could be transformed into the two-activities sub-sequence $login - request\ help$.

1.4 Objectives of the research

Below we present the main objectives of this research along with their particular objectives:

1. To obtain a set of contrast patterns that characterises different classes of visitors based on conversions and that is suitable for the interpretation of a Marketing expert.

    1.1 To create classes of visitors according to the conversions that they performed on the website.

    1.2 To obtain a set of descriptive contrast patterns for the visitors from each class.

2. To obtain a graphic representation that describes, at a high-level, the navigation behaviour of visitors, and that is suitable for the interpretation of a Marketing expert.

    2.1 To identify sequences of visited pages of frequent occurrence in sessions that belong to the same class of visitors.

    2.2 To create a graph that represents the navigation behaviour of a class of visitors using the most frequent sequences of visited pages.

    2.3 To verify if the previous representation allows us to compare different classes of visitors.

We followed a methodology that allowed us to achieve the objectives of the research. That is described in the following section.

1.5 Methodology

In this research, we achieved two main objectives following the methodology shown in Figure 1.1. For the first objective (which is the characterisation of visitors according to the conversions they make) we followed a four-step methodology. It consists of identifying visitors, extracting features and filtering bots, create classes of visitors according to conversions, and pattern mining. It is shown inside the blue square in Figure 1.1. For the second objective (which is to describe the navigation behaviour of a whole class of visitors), the methodology was three-step. It consists of extracting the representative sequence of pages (rules), representing sessions with rules, and visualizing sessions represented with rules. It is shown inside the orange square in Figure 1.1. Next, we explain both methods in more detail.

Characterisation of visitors according to conversions

To characterise the website visitors according to the conversions they make, we obtained a set of distinguishing characteristics (patterns) for each type of visitor (class). From the input data, which are web log files, we identified the activities performed by each unique visitor in a period. That is called a session. To obtain descriptive patterns, one has to provide descriptive characteristics. Therefore, we extracted features that describe sessions. Sessions could be performed by human

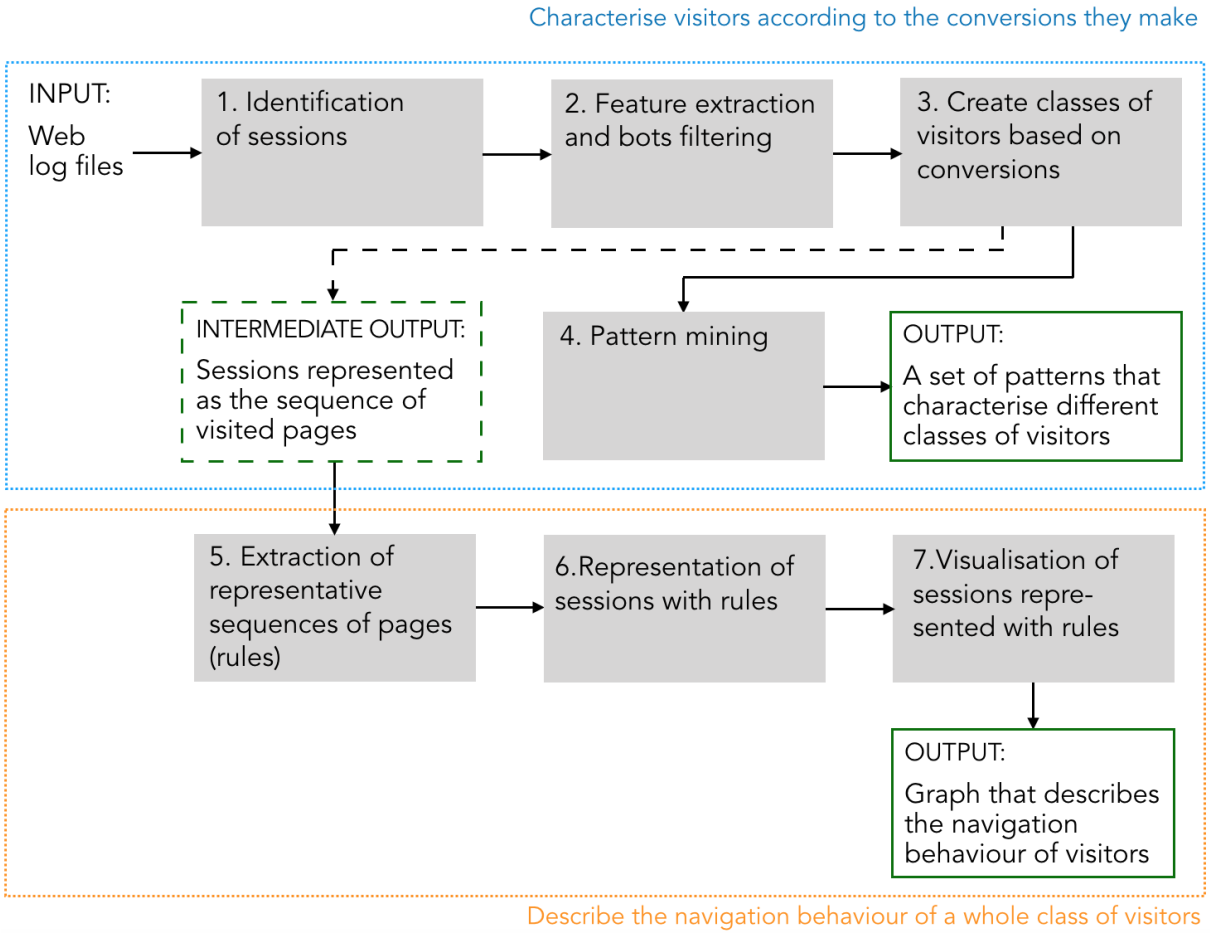

Figure 1.1: The methodology of this research.
Inside the blue square, we show the four-step method for characterising visitors according to the conversions they make. Inside the orange square, we show the three-step method for describing the navigation behaviour of visitors.

visitors or by bots. However, we are only interested in characterising human sessions. Therefore, we removed bots using a machine learning algorithm. Having the set of human sessions, we created classes of visitors based on the conversions performed in each session (session labelling). The above has resulted in two outputs:

1. The human sessions represented as descriptive features. We used this output to mine patterns, and obtain our first output: **a set of patterns that characterises each class of visitors.** Patterns provide the characteristics that are found in a certain percentage of visitors. That allows focusing on the characteristics that are relevant among visitors from a given

class.

2. The human sessions represented as the sequence of visited pages. This output was used as input data to describe the navigation behaviour of visitors. Next, we explain this process.

Description of the navigation behaviour of visitors

To describe the navigation behaviour of visitors, we used the human sessions represented as the sequence of visited pages. In a given class of visitors, we extracted the most frequent sequences of visited pages. We called those frequent sequences "rules". Each rule is formed by different pages and represents visitors actions. For example, making a payment or searching for the availability of a product. We named the rules and used them to obtain a reduced representation of each session. The representation of sessions with rules drastically reduced the length of sessions (for example, from forty pages to two rules). We used all the reduced sessions of the class to create a visualisation. That is the second output: **a graph that describes the navigation behaviour of visitors**. This graph shows the distinguishing behaviour of the visitors from the same class. The graph also eases to compare the navigation behaviour of different classes of visitors. It prevents the Marketing expert from analysing huge graphs for understanding the navigation behaviour of visitors.

In the next Chapters, we deeply explain our motivation and methodology to achieve the objectives of this research. In Chapter 2, we present the limitations of existing Web Analytics Solutions. In Chapter 3, we explain in more detail how we characterised the website visitors. Finally, in Chapter 4, we describe our method for describing the navigation behaviour of visitors.

Chapter 2

Previous work

To improve the effectiveness of marketing strategies, it is valuable to know the characteristics of the visitors who perform a conversion and understand their navigation behaviour. This knowledge has to be extracted from the huge amount of data generated to monitor the website traffic. This falls on the web analytics field, which is concerned with measuring, analysing, and reporting website data to obtain knowledge about website usage [22]. Current Web Analytics Solutions (WAS) provide valuable information about the webpages, visitors and marketing campaigns. Nevertheless, they have some limitations for characterising website visitors and providing a high-level description of their navigation behaviour.

The characterisation of visitors is the identification of the properties that describe a class of visitors and distinguish it from other classes. Web analytics software (like Google Analytics) does not provide a characterisation as such. Instead, it provides tables and charts to analyse characteristics independently. Most machine Learning approaches found in the literature use supervised learning to filter bots or use unsupervised learning to find clusters of visitors. However, those approaches do not address the characterisation of different classes visitors for marketing purposes. To fill in that gap, in this research, we characterise different classes of visitors according to conversions. The use of conversions as classes of visitors ease the interpretation of results for marketing experts.

The description of the navigation behaviour of visitors consists of extracting the most relevant actions performed by visitors from a given class. Existing web analytics software provides a

page-by-page detail of the navigation of visitors. However, there is usually a huge amount of visitors and possible navigation paths. Thus, the navigation of a whole class of visitors would result in tens of pages to visualise. Web log mining techniques are used to cluster visitors, but they do not provide a high-level description, either. Due to this lack, we propose a method for obtaining a high-level description of the navigation behaviour of visitors. It drastically reduces the amount of data that a marketing expert has to analyse for understanding how visitors navigate a website.

In this chapter, we present existing solutions related to this research. In section 2.1, we describe the previous work related to the characterisation of visitors. Then, in section 2.2, we present previous approaches to describe the navigation behaviour of visitors. Finally, in section 2.3, we present our conclusions. Throughout this chapter, we describe commercial software and literature research. From the wide variety of commercial software, we centre our analysis on Google Analytics and Matomo. Google Analytics is the most popular web analytics software [105, 9, 1]. Matomo, on the other hand, is an alternative to some limitations of Google Analytics [5], and it was used in a previous approach to this research [38]. Both provide similar functionality.

2.1 Characterisation of visitors

The characterisation of visitors is the identification of the properties that describe a class of visitors and distinguish it from other classes [48, 52]. Knowing the properties of visitors who perform a given conversion allows designing more effective strategies. For example, marketing strategies could be directed to visitors with similar characteristics [83]. However, existing Web Analytics Solutions have some limitations for characterising website visitors. Below we explain those limitations. In Subsection 2.1.1, we explain the limitations of popular commercial software. In Subsection 2.1.2, we describe the limitations of literature approaches.

2.1.1 Commercial software for characterising visitors

Google Analytics and Matomo have similar web analytics reports, but none of them provides a characterisation of visitors. Instead, they show scattered tables and charts to analyse the characteristics of visitors. These are grouped in different categories, for example, demographics, interests, and geographic. In Figure 2.1, we show an example of the charts shown in the category "demographics", in Google Analytics. Characteristics can be used to create segments, and different segments can be compared [2, 5]. A segment is a portion of visitors whit certain characteristics, for example, "male visitors from México who are between 25 and 54 years old". However, the main disadvantage is that the characteristics of segments are shown separately. There is not a report that sums up the distinguishing characteristics for a segment of visitors.

To characterise website visitors, we used conversions associated with a five-step sales funnel. Sales funnels can be built with Google Analytics or Matomo [2, 5]. However, they do not provide a characterisation of visitors using common steps of the sales funnel (e.g. awareness, consideration, intent, purchase and loyalty). Instead, a sales funnel built with Google Analytics or Matomo shows the sequence of pages to reach a given goal (conversion). For example, making a purchase or submitting a form [2, 5]. These sales funnels provide information about the visited pages but do not show the characteristics of visitors at each step. In Figure 2.2, we show an example of a sales funnel created in Google Analytics.

2.1.2 Literature approaches for characterising visitors

There are different web log mining approaches in which the characteristics of visitors are analysed. However, they are centred around identifying clusters of visitors for different purposes, not on characterising visitors. For example, J. J. Prabhu et al. [47] propose to group website visitors for prioritizing live, human-assisted, support. They use different features to group visitors on-line, for example, browser, country, current page, campaign source, landing page, etc. They prioritise live support based on the group to which the visitor belongs.

Another example is from L. Maxwell et al. [54]. They use website data to group visitors and determine malicious sources. They use variables like country, browser, a hardware identifier, and

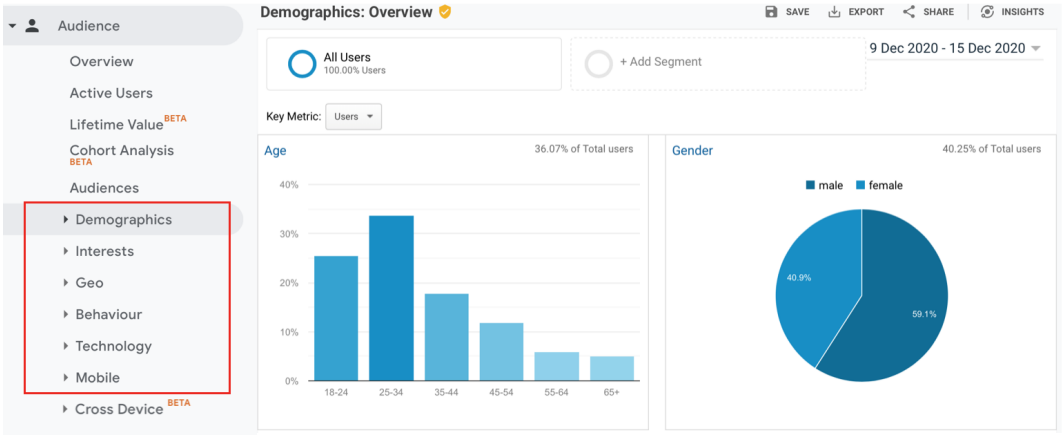

Figure 2.1: Example of charts in the category “Demographics” in Google Analytics. Each tab in the red square shows different tables or charts. Age and gender charts correspond to the tab “Demographics”. The period is modified on the top right.

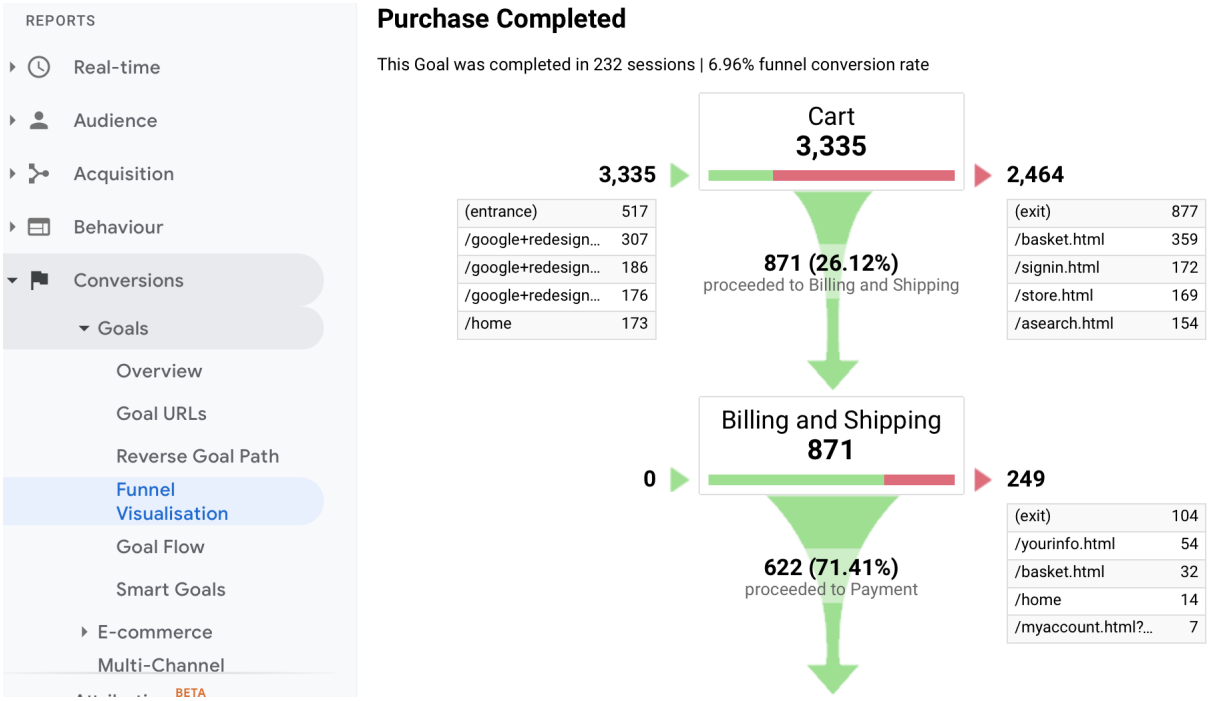

Figure 2.2: Example of a sales funnel created in Google Analytics. In Google Analytics, a funnel shows the sequence of pages to reach a given conversion goal. In this example, we see the first two pages of the conversion goal “Purchase Completed”. Those pages are “Cart” and “Billing and Shipping”. On the left, they show the previously visited page. On the right, they show the exit page of visitors leaving the website.

the presence of cookies. They determine possible combinations of variables in pairs, for example, "Country-URL" and "Browser-URL". They create vectors to capture the distribution of each pair of variables. Then, they execute one or more clustering algorithms on the vectors (for example, k-means or hierarchical clustering). Data is compared with known normal and abnormal data. When a cluster contains a significant representation of a pair of variables known to be normal or abnormal, it is considered that all visitors from the cluster are of the same type.

An example of unsupervised learning is from M. Santhanakumar et al. [58]. They calculate user similarity and session similarity using Euclidean distance. User similarity is obtained with the visitor attributes (e.g. the IP address and visited pages). Session similarity compares two sessions. They use the user and session similarities to find clusters. The authors state that this may be useful to improve information accessibility on websites.

The visitor attributes have also been used to improve website performance or navigation experience. For example, Scott et al. [103] use the visitor attributes for routing webpage traffic. They obtain the visitor attributes and previous performance of each webpage. They use this information to predict the performance that a visitor would have on each webpage variant. Based on that prediction, they direct the visitor to a specific webpage. This approach allows for improving website performance.

There is extensive Machine Learning research to use the characteristics of visitors for filtering bots [76, 21, 38, 39, 95]. However, they do not address the characterisation of visitors for marketing purposes. We concluded that the characterisation of visitors using conversions as classes has not been previously addressed. In the next subsection, we present the previous work related to the description of the navigation behaviour of visitors, which is our second objective in this research.

2.2 Description of the navigation behaviour of visitors

The objective of describing the navigation behaviour of visitors is to understand how visitors navigate a website. However, a huge amount of data has to be analysed for understanding the navigation behaviour of a whole class of visitors. Therefore, an aggregated high-level description

would be valuable. Existing Web Analytics Solutions have some limitations to provide such a description. In this section, we explain those limitations. In Subsection 2.2.1, we explain the limitations of popular commercial software. In Subsection 2.2.2, we describe the limitations of the literature approaches.

2.2.1 Commercial software for describing the navigation behaviour

Google Analytics and Matomo provide similar functionality for tracking the navigation behaviour of visitors. In Google Analytics, it is called "Behaviour flow" report. In Matomo, it is the section "Goal conversion tracking", but it is only available in the premium version. Both consist of showing the sequence of most visited pages in a period. It is aimed to measure the engagement page to page. Therefore, it is useful for finding pages where the traffic is lost, but it is difficult to follow a path with numerous pages. It is also difficult to visualise the path of 100% of visitors if they are numerous and behave differently. It is possible to track events instead of pages. Nevertheless, those events have to be previously configured. Therefore, events do not represent the natural navigation behaviour of visitors. In Figure 2.3, we show an example of the behaviour flow chart in Google Analytics.

Figure 2.3: Example of the Behaviour Flow report in Google Analytics.
It shows the sequence of most visited pages, from left to right. The number of pages (interactions) can be incremented as needed. The thick red lines indicate traffic drop-offs.

2.2.2 Literature approaches for describing the navigation behaviour

Web Log Mining or Web Usage Mining is the utilization of different data mining techniques for obtaining knowledge about the navigation behaviour of visitors [45]. There are diverse web log mining approaches in the literature. However, they are centred around identifying clusters of visitors not on obtaining a high-level description of their navigation behaviour. The analysis of visited pages, called sequence mining, is commonly applied for discovering patterns with a frequency support measure [37]. The most common sequence mining approach is clickstream analysis for clustering visitors [42]. Clickstream is the sequence of pages visited by a user in a given website and period [19]. Below we describe some clickstream approaches.

S. Tiwari et al. [94] use previously visited pages to forecast online navigational pattern (finding next page expectation). They apply agglomerative clustering to group visitors according to the previous web data accessed. They obtain the set of frequently visited pages in each group of visitors. This information is used to put in the cache pages with higher frequency, to reduce the search time.

A. Banerjee et al. [19] propose to find the longest common subsequence of clickstreams using a dynamic programming algorithm. Then, they identify similar users by computing a similarity value that considers the time spent on each page. With the similarity values, they construct a weighted similarity graph. Finally, they find clusters on that graph. They found that, in some cases, there are no exact matches. As a solution, they propose to first group data into categories.

There are other clickstream pattern mining approaches, but they are focused on improving the runtime or memory consumption for clustering visitors [42, 41]. We found no approaches that use clickstream analysis for providing a high-level description of the navigation behaviour of visitors. Visualisation tools have also been proposed to analyse the navigation behaviour of visitors [38, 21, 23, 80, 17, 61]. However, they provide solutions for a detailed analysis of web pages. For example, to find the percentage of visitors on each web page. But they do not address the creation of a high-level representation of the navigation behaviour of visitors.

2.3 Conclusions

In this chapter, we presented previous work related to the characterisation of visitors and the description of their navigation behaviour. We addressed the limitations of both commercial software and literature approaches.

Regarding the characterisation of visitors, commercial software shows metrics in multiple tables and charts [2, 5]. Therefore, it is not easy to find multiple-feature patterns or group visitors into classes that are not previously configured. Another disadvantage of commercial software is that configuration changes do not apply retroactively [2, 5]. For example, goals and conversions will be calculated using data from the time of the change onward, but do not affect historical data.

Literature approaches have overcome some limitations of commercial software for analysing the characteristics of visitors. Machine Learning algorithms have been used for finding clusters of visitors with different purposes [98, 47, 54, 58]. For example, identify bots [76, 21], prioritise live support [47], detect the learning style of students in online platforms [69], or identify malicious sources of website traffic [54]. However, the literature approaches have not addressed the characterisation of visitors using conversion-based classes.

From the above, one contribution of this research is to characterise different classes of visitors using classes based on conversions. We achieved it by obtaining a set of contrast patterns for the visitors who perform each type of conversion. With this method, we obtained the distinguishing characteristics of each class of visitors. Besides, the use of conversions as classes eases the interpretation of results for marketing experts.

Regarding the description of the navigation behaviour of visitors, commercial software provides a page-by-page detail of visited pages, but it does not provide a high-level description of the navigation behaviour [2, 5]. Tens of pages would have to be reviewed to understand the navigation behaviour of a whole class of visitors. Due to the limitations of the commercial software, clickstream analysis has also been applied for understanding how visitors navigate a website [78, 52, 51, 44, 94, 19]. However, clickstream analysis approaches are focused on finding clusters of visitors [94, 19, 42, 41], not on obtaining a high-level description of the navigation behaviour of visitors.

The previous limitation motivated our second contribution, which is to provide a high-level description of the navigation behaviour of visitors. A distinguishing characteristic of our approach is that we extract the real navigation behaviour of visitors, instead of finding how visitors behave in previously known actions. It is important because there is usually a lot of difference between expected and real navigation behaviour [28].

In the next Chapters, we describe how we achieved the two contributions of this research. In Chapter 3, we explain how we characterised visitors using classes based on conversions. Then, in Chapter 4, we describe how we obtained a high-level description of the navigation behaviour of visitors. Finally, in Chapter 5, we present our conclusions and future work.

Chapter 3

Characterisation of visitors

One way for measuring the value of a website is by its effectiveness in getting visitors to perform conversions. In this chapter, we aim to characterise visitors in terms of the conversions they make, and how types of visitors contrast one another. This is valuable to confection a strategy for encouraging certain kinds of visitors to make a conversion that otherwise would not.

To achieve this goal, one needs to set apart what counts as a visitor to a website in the first place. A visitor makes explicit reference to a unique "user", browsing over the website and during a certain time. The collection of all the user activities in relation to one or more conversions is called a session. Hence, a visitor is a collection of sessions occurring during a period.

Given a web log recording activity in a website during a period, one can identify the set of all visitors to the site. Nevertheless, for a correct analysis of the performance of the website, one has to extract the features (raw or composed) that best describe visitors and remove any activity that suspiciously comes from a non-genuine visitor. To extract adequate features, we conducted feature engineering and obtained a feature space that describes sessions. To identify non-genuine visitors, we first used the descriptive features that we obtained, to identify what counts as genuine behaviour and not. Building upon previous research, we then applied different classification mechanisms, which helped us separate lots of bots even on supposedly clean web logs, according to our partner security experts.

Each session now in our dataset is allegedly genuine and can be easily associated with one or more conversions. Conversions allow converting prospective browsers into loyal buyers. But that

is reached by multiple steps, known in the marketing field as the sales funnel. To design and improve conversions, marketing experts have to understand the characteristics of visitors at each step of the sales funnel. Therefore, besides associating each session with conversions, we have partitioned all types of conversion into a five-layer sales funnel.

Having sessions associated with conversions, we have considered each type of conversion as a class of visitor to characterise. To this aim, we have used a miner based on contrast patterns. So, each pattern is so that it characterises what elements of a class have in common and how one class differs from the others. We have successfully identified a set of patterns for each class. We manually interpreted these patterns and obtained relevant insights for each step of the sales funnel. Those results have been validated by our partner company, from both departments of Information Technology and Marketing.

In this Chapter, in section 3.1, we explain how we identified visitors. In section 3.2, we describe how we obtained descriptive features for each session and removed bots. In Section 3.3, we describe how we associated sessions with conversions and the sales funnel. In Section 3.4, we explain how we mined and interpreted contrast patterns. Finally, in Section 5.1, we present our conclusions.

3.1 Identification of visitors

A visitor is the collection of sessions that occurred at a certain time and correspond to the same unique “user”. One contribution of our work is to carry out an analysis that is oriented to classes of visitors in terms of sessions. This requires an analysis of visitor behaviour as of an entire session, each of which could be then labelled. The initial data were logs from a commercial website. Therefore, the first step of our methodology is to identify visitors in terms of sessions. To this aim, we followed a two-step process that is described in this section. In the first step, we transformed each web log entry into an object with descriptive information (e.g. city, hour, and downloaded bytes). In the second step, we grouped the output of the previous step into a sequence of sessions for each visitor.

3.1.1 Obtaining descriptive objects from web log entries

The input data were web logs that contained 3 million requests. Those logs correspond to two weeks of web traffic. Each **request** represents a web page petitioned to the server. Logs had an extended version of the NSCA Common Log Format (CLF). Figure 3.1 shows an example of the CLF. Below we briefly describe each part of it, according to information published on the Apache website and HTTP specification [4, 3].

1. **IP address:** it is the IP address used to make the request.
2. **RFC:** it is usually "-". It indicates that the information is not available.
3. **User ID:** it is an identifier of the person who requests the document. It is usually "-", unless the document is password protected.
4. **Date time:** it indicates when the request was received.
5. **The request line from the client:** it contains the method used to make the request (GET, HEAD, POST), the resource requested (URL), and the protocol used.
6. **Status code:** it is the status that the server sends to the client. There are five status categories: a successful response (codes 2xx), a redirection (codes 3xx), an error caused by the client (codes 4xx), or an error in the server (codes 5xx).
7. **Size:** the size (bytes) of the content sent to the client. It does not include the response headers.
8. **Referrer:** it indicates from where the client was referred.
9. **User-agent:** it is information about the browser used by the client.

We transformed each web log request into an object describing the request, but with information that is oriented for human consumption. For example, instead of recording the visitor user-agent, we obtained the web browser, operating system, and device. In Table 3.1, we indicate the features that we obtained from the Combined Log Format (CLF). Whenever we mention **request features**,

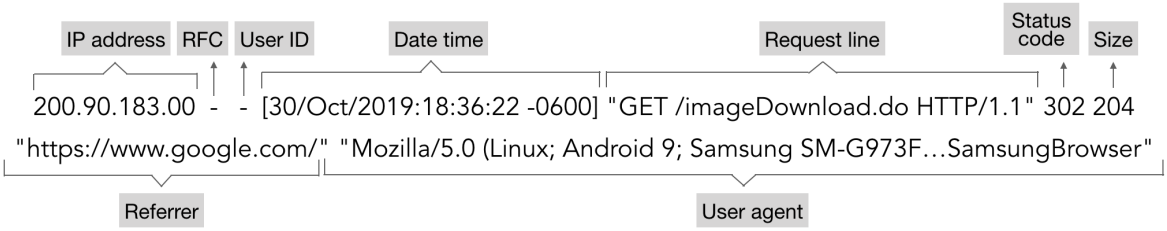

Figure 3.1: Structure of the Combined Log Format.

we refer to the features listed in the second column of Table 3.1. In the next Subsection, 3.1.2, we explain how we used the request features to identify the sessions associated with each visitor.

Table 3.1: Features extracted from the Combined Log Format (CLF).
The description of each feature is in Appendix A. "Not applicable" means that no feature was extracted from this part of the CLF.

Part of the CLF	**Request features (Features extracted from the CLF)**
IP address	IP address, City, Country, subdivision, Organization.
RFC	*Not applicable*.
User ID	*Not applicable*.
Date time	Date time, Day of the week, Hour.
Request line	Method, Requested URL, URL type, Depth.
Status code	Status.
Size	Bytes downloaded.
Referrer	Referrer URL.
User-agent	Web browser, Operating system, Device.

3.1.2 Identifying sessions of each visitor

The process of identifying the sessions of each visitor is two-step. In the first step, we identified the requests that belong to the same visitor using a **fingerprint**. In the second step, we grouped

the requests of each visitor into sessions.

Identification of requests per visitor

In previous research [97, 31, 101], the identification of visitors was obtained using different combinations of the IP address, the browser, the operating system (OS), the browser version, and the OS version. We composed the fingerprint with *IP address + device name + OS name*. We omitted the browser and OS versions to avoid identifying a visitor as different upon a system update. This fingerprint does not allow us to identify multiple devices of the same user, but, due to privacy reasons, we rejected the use of other tracking tools (like web beacons or script codes). Using the described fingerprint, we identified 24,000 different visitors.

Grouping of requests as sessions

Using the fingerprint, we extracted two datasets: 1) a list of visitors that, in later steps, is used for identifying if a visitor is new or recurrent and 2) the list of requests executed by a visitor, chronologically ordered. The list of requests describes visitor actions from one or more sessions. Having identified the first visitor action (request) on a session, we marked the end of that session upon the occurrence of a visitor action that leads to a 25-minute interval of visitor silence (no activity). Then, the next visitor action, if any, becomes the first entry of a new session. The 25-minute interval has been commonly applied in previous web traffic analyses [97, 31] and it is similar to the interval used by Web Analytics Software [2, 5].

The result was a list of 66,000 sessions grouping their correspondent requests. The next step was to obtain the features that summarise each session (e.g. the duration of the session from the first to the last request). That is described in the next section, 3.2.

3.2 Feature extraction and bot filtering

To obtain a descriptive and reliable characterisation of visitors, it was necessary to extract the features that best describe sessions and remove sessions suspiciously performed by a non-genuine

visitor. Those steps are described in this section.

3.2.1 Feature extraction

Taking advantage of previous approaches from the Machine Learning Group [38, 21, 76], we analysed the feature space that was used. We also analysed the features commonly extracted in other web log mining research [101, 97, 31, 16, 86, 79, 59, 43] and the features generally used by Web Analytics Software [2, 5]. As a result, we proposed the feature space for describing the sessions of any website. Below we present that feature space grouped in seven categories. We only mention the name of the features that belong to each category and some clarifications. A description of each feature is in Appendix B. Two asterisks (**) indicate a feature that had not been used in previous approaches to this project and one asterisk (*) indicates a feature that was obtained in a previous approach using Matomo [5].

1. Session timing: *Day of the week*, *Hour*, *Session duration**, and *Average time per page***. This category groups the features related to the moment in which the session happened:

2. Website navigation: *Depth of visited pages*, *Number of requests per session***, *Known referrer***, *Referrer URL*, *Entry page*** and *Exit page***. This category provides information about the navigation behaviour in each session.

3. Geographic: *City*, *Country*, *Subdivision*, and *Organization*. This category locates the session based on the IP address and using geolocation databases. Previous approaches [38], [76] also used ASN, ASN name, country code, ISP, latitude, longitude, region code, time zone, zip code, and continent code. We did not include them because we found them redundant.

4. Type of visitor: *New visitor** and *Known bot*. This category indicates 1) if the visitor is new, and 2) if the visitor is a known bot. *New visitor* is set to "True" if the fingerprint of a given session is not in the list of visitors obtained in 3.1.2. In that case, the fingerprint is added to the list of visitors. *Known bot* is set to "True" if the visitor appears in public databases in which known bots are registered. Sessions from known bots are eliminated, as explained in Section 3.2.2.

5. Hardware and software: *Web browser*, *Operating system* and *Device*. This category describes the hardware and software used to navigate the website. We extracted them without version because it is not relevant for characterising visitors.

6. Size and format of visited pages: *Bytes downloaded per session***, *Requested robots file***, *Percentage of image files***, *Percentage of java server pages***, *Percentage of HTML files***, and *Percentage of pdf files***. Sessions that requested the robots file are considered bots and are eliminated as explained in Section 3.2.2.

7. Request method and server response: *GET method percentage***, *HEAD method percentage***, *POST method percentage***, *Percentage of informational status (1xx)***, *Percentage of success status (2xx)***, *Percentage of redirection status (3xx)***, *Percentage of client error status (4xx)***, and *Percentage of server error status (5xx)***. This category is targeted at Information Technology experts. It could allow identifying unusual navigation behaviour and website errors that may produce a bad user experience or low conversion rate, e.g., a high percentage of server error status may indicate a broken link in a web page, and a high percentage of the HEAD method may indicate the presence of bots.

We extracted the described features, obtaining a dataset that describes sessions as feature vectors. Each session also has the sub-list of the requests that are part of the session. The next step of the methodology was to remove non-genuine visitors, known as bots.

3.2.2 Bot filtering

At this point, we have sessions represented with descriptive features. Nevertheless, bots can be up to 37.9% of website traffic [30]. To obtain correct conclusions, based on the characteristics of human visitors, one has to filter bots. Therefore, we filtered bots with two methods: 1) a fast-filter and 2) a machine learning model. The fast-filter is based on the binary features “Known bot” and “Requested robots file”. The first feature was used to identify the sessions that come from known-bots sources. The second feature indicates if the robots file was requested. We eliminated sessions with the value “True” in at least one of those two features. In the machine learning

model, we used the method previously used by the Machine Learning Group. That method uses Bagging-RandomMiner as an anomaly detection algorithm. It allows identifying atypical values in the dataset. Bagging-RandomMiner has shown good results in data mining tasks [50], and the Machine Learning Group has specifically tested it for filtering bots [21].

Using the previous methods, we found 23% of bot traffic in the dataset. That percentage is high, considering that the input data was supposedly clean of bots, according to our partner security experts.

After removing bots, data was ready to be partitioned into different classes of visitors. We explain this process in the next section, 3.3.

3.3 Session labelling based on website conversions

One contribution of our approach is the use of conversions, associated with the sales funnel, as classes of visitors. To this aim, we followed a two-step process. In the first step, we associated the conversions that a visitor can perform on the website with the sales funnel steps. In the second step, we used conversions for labelling data. Below we explain these steps.

3.3.1 Classification of conversions in the sales funnel steps

We started by identifying web pages that correspond to different types of conversions (e.g. request online help or make a payment). We obtained at least one URL per conversion. Then, we manually associate each conversion with the corresponding step of the sales funnel. We used a five-step sales funnel whose steps are: 1) Awareness, 2) Consideration, 3) Intent, 4) Purchase, and 5) Loyalty. The Figure 3.2 shows the identified conversions associated with the sales funnel steps. Below we describe each step.

1. Awareness: it is the first step of the sales funnel whose objective is to be discovered by potential clients. In this step, it is expected the highest number of potential clients. Strategies at this step aim to brand or product presence. The awareness conversions that we identified on the

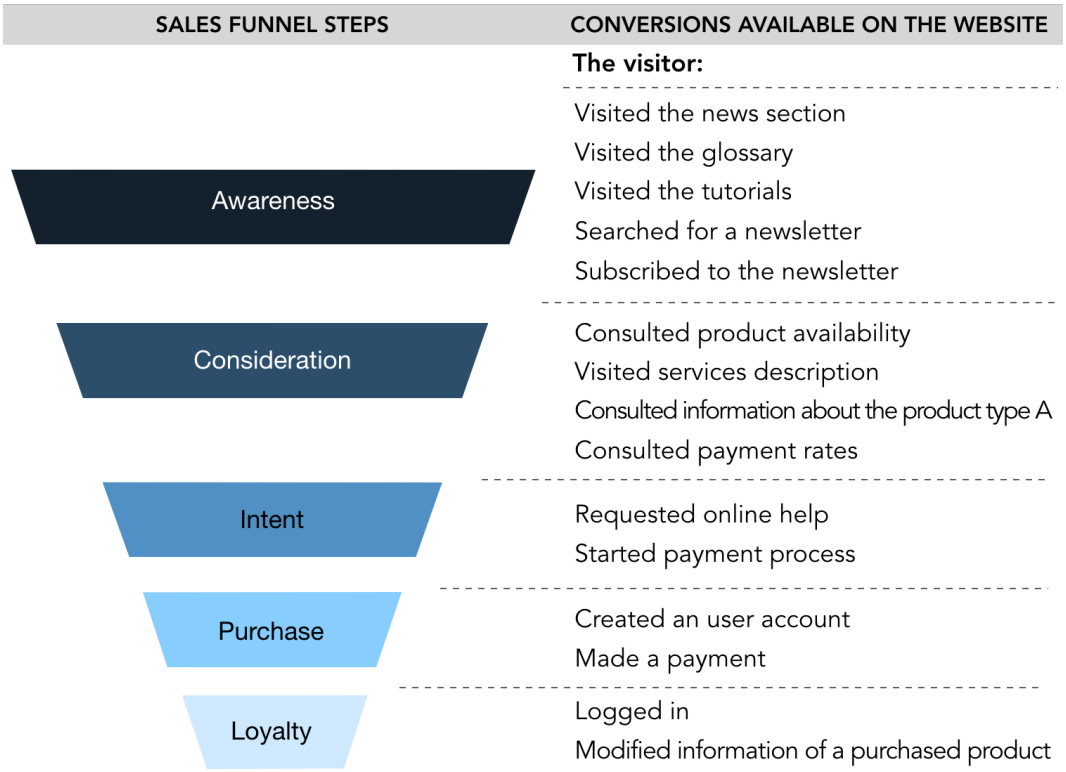

Figure 3.2: Conversions that are available on the analysed website.
Each conversion listed on the right corresponds to one or more pages of the website (URL), and it is associated with the sales funnel step shown on the left. The leaky funnel represents the dynamic of gradually losing clients (or potential clients) that become uninterested at some point.

website are a page with news about the industry, a glossary, open access tutorials, and a newsletter to which visitors can subscribe [46, 91, 100].

2. Consideration: at this step, potential clients start to evaluate different options of products that could fulfil a latent necessity or desire, without necessarily having the intention of making a purchase. It is relevant to provide enough information about the value and advantages of the brand or product. The consideration strategies that we identified are a page for verifying the availability of products, pages with the description and rates of products and services, and pages related to a product that the website strongly promotes, which we called “product type A” [46, 91, 100].

3. Intent: this is the step in which a potential client chooses the specific product to buy and is eager to pay for it. Strategies should ease and support the purchase process. Any doubt or

problem that is not nimbly solved could lead to shopping cart abandonment. The intent strategies that we identified on the website are online help available (on request, not proactive) and the pages that are part of the payment process. The payment confirmation page is not associated with this step. We found no strategies for avoiding cart abandonment, e.g. buttons or pop-ups if the visitor moves away from the purchase process [46, 91, 100].

4. Purchase: it refers to the moment in which a potential client pays and becomes a client. The perception of buyers about the buying experience and the quality of the product will impact the probability of buying back or recommending the product. The purchase strategies that we identified on the website are the pages related to the payment confirmation and the creation of a user account (because this account is a requirement for paying). We found no strategies to promote buying back or evaluating the purchase experience after paying [46, 91, 100].

5. Loyalty: it refers to get clients buying back or promoting the brand or product. We found no aggressive loyalty strategies on the website(e.g. rewards o referrers program). We placed as loyalty conversions actions that can be performed by someone who has already bought: logged visitor and modify information of products that were previously bought [46, 91, 100].

3.3.2 Use of conversions for labelling data

To classify visitors based on the conversions that they performed, we used each conversion listed in Figure 3.2 as a binary feature. We identified conversions in two steps. In the first step, for each **request**, we determined the conversion performed, if any. In the second step, we identified the conversions performed in each **session** based on the conversions of the requests that are part of the session. For each session, we used another binary feature called *conversion*. It was set to "True" if at least one of the 15 possible conversions was performed in the session.

The 16 features were used as a label. This way, we classified sessions according to conversions associated with the sales funnel. The labelled dataset was the input data for two tasks:

- For mining contrast patterns as described in Section 3.4.

- For describing visitors using sequence mining, as described in Chapter 4.

In Figure 3.3 we summarise the features extracted from each session, including conversions used as labels.

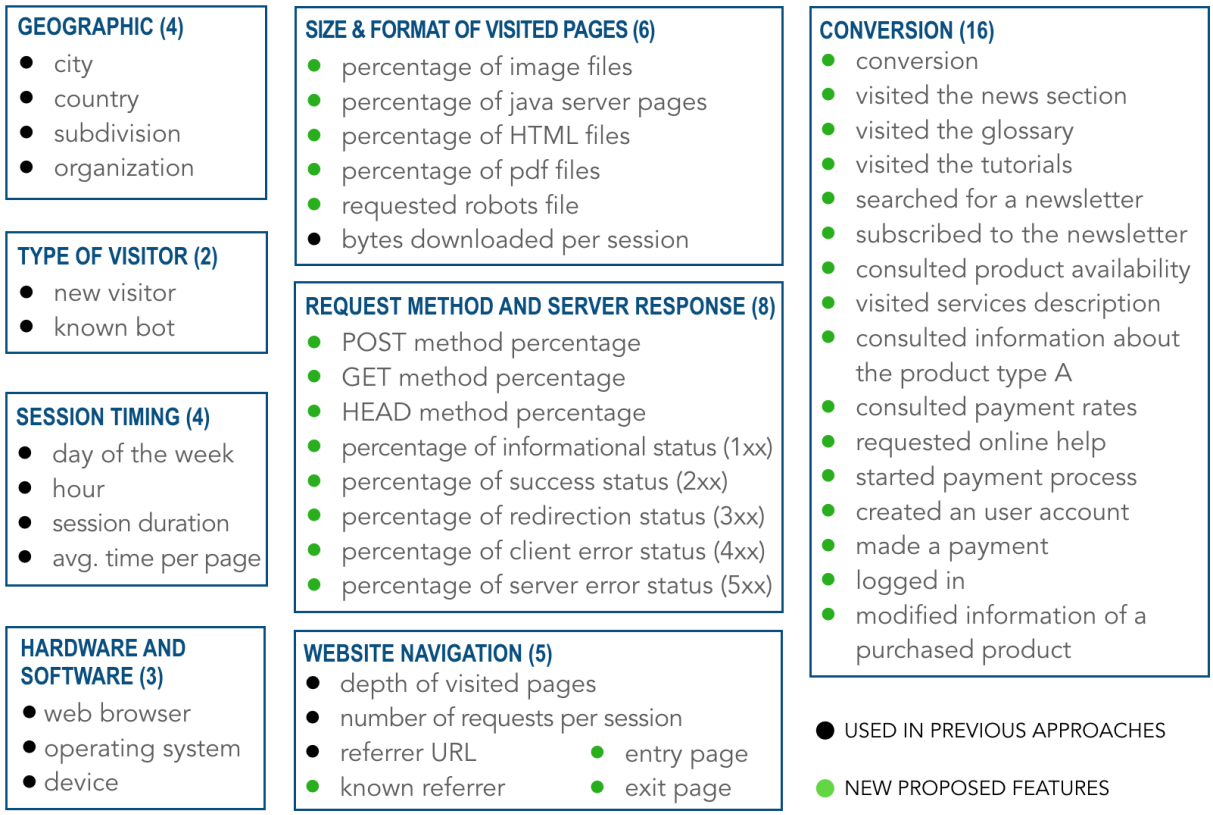

Figure 3.3: Summary of features extracted from sessions.
Features with a black dot were already used in previous approaches of the Machine Learning Group at Tecnológico de Monterrey. Features with a green dot were proposed in this research.

3.3.3 Impact of bots on each conversion

Taking advantage of the conversions used as labels, we measured the impact of bots on each conversion. Consider C a given conversion associated with the sales funnel, T_C the total number of sessions that performed C, and B_C the number of sessions that performed C and were identified as bot traffic. We calculated the impact of bots on each conversion C, denoted BI_C with the formulae $BI_C = \frac{B_C}{T_C}$ expressed as a percentage. In Figure 3.4, we show the impact of bots on each conversion.

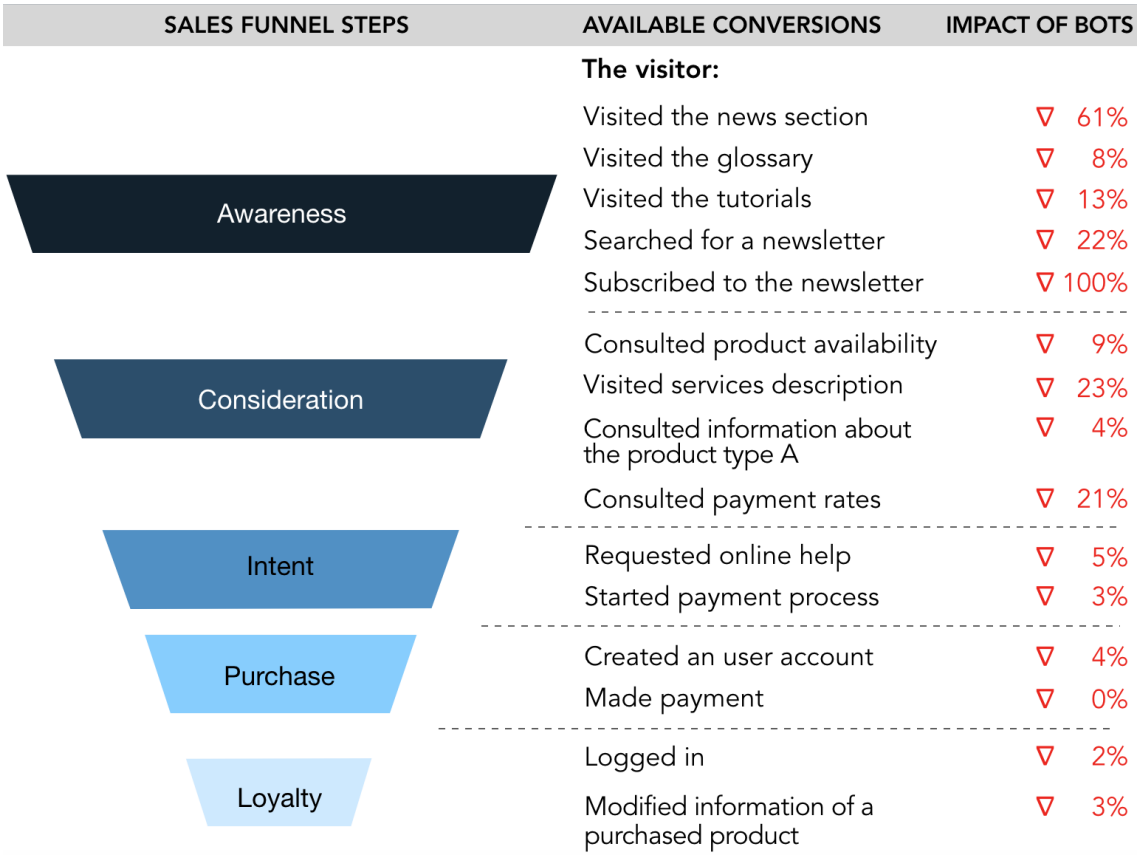

Figure 3.4: Impact of bots on each conversion.
In the middle, we list the conversions that a visitor could perform on the website. Each feature is associated with the sales funnel step at the left. At the right, we indicate the percentage of sessions identified as bots, e.g., from the visitors who visited the news section, 61% were bots.

According to the feedback of our marketing partner, knowing the impact of bots on each type of conversion was relevant, especially because we found a high impact of bots traffic on specific conversions. For example, 100% of visitors who subscribed to the newsletter were bots.

After labelling data and measuring the impact of bots on each conversion, we applied data mining techniques to find what the different classes of visitors have or do not have in common. That is explained in the next section, 3.4.

3.4 Pattern mining

To characterise visitors, we were interested in finding out the characteristics of the visitors who made each conversion and the visitors who did not. Therefore, we applied data mining techniques based on contrast patterns. Part of this process is to select the algorithm to use for mining patterns. That selection depends, among other aspects, on the characteristics of the dataset to use. Therefore, in this section, we explain the pattern mining process in four steps: dataset exploration (3.4.1), selection of the contrast pattern algorithm (3.4.2), implementation of the selected algorithm (3.4.3), and interpretation of patterns (3.4.5).

3.4.1 Dataset exploration

We explored two aspects of the dataset: the percentage of visitors that belong to each class and the feature values. Below we explain how we did it.

Exploration of classes to contrast

As explained in Section 3.3, we labelled data using 16 binary features. 15 features indicate the accomplishment of specific conversions (e.g. if the visitor subscribed to the newsletter or not) and one feature indicates if the visitor performed at least one conversion or not. We computed the proportion of sessions that performed each conversion. We found that 42% of the visitors performed at least one conversion and, in Figure 3.5, we list the percentage of sessions that performed each one of the 15 conversions. After filtering bots, there were no visitors who subscribed to the newsletter. Therefore, we did not use this conversion further.

In Figure 3.5, we can see that the percentage of visitors who performed each type of conversion is too low in some cases. It means that we are dealing with imbalance classes. We will consider this characteristic of the dataset for selecting the contrast pattern mining algorithm.

The second aspect to explore in the dataset was the feature values. That is explained next.

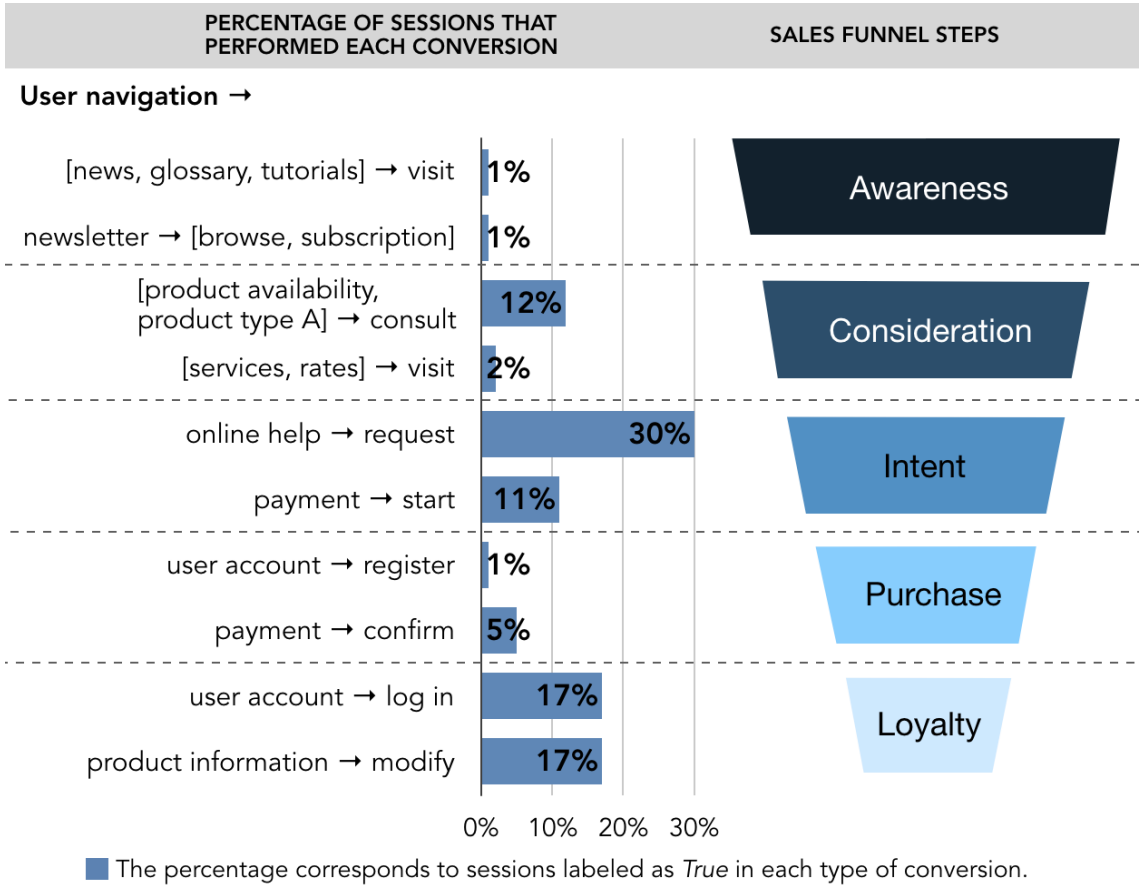

Figure 3.5: Percentage of visitors who performed each type of conversion on the website. Conversions listed at the left are binary features used as the class for mining contrast patterns. Each feature is associated with the sales funnel step at the right. The bars in the centre, indicate the percentage of visitors who performed each conversion.

Exploration of feature values

Features of the categories *Request and server response* and *Size and format of visited pages* are aimed to highlight failures or risks. They are relevant only if they are suspiciously high or low. For example, the "GET" method indicates that the visitor requested all the content of the page. Most traffic is expected to use this method. Hence, it is not relevant if it is found in 90% of visitors. On the contrary, the "HEAD" method is commonly used by bots. It is useful only if we find a high percentage. After exploring the values of the dataset, we eliminated features of the mentioned categories because they were not suspiciously high or low.

3.4.2 Selection of the pattern mining algorithm

After exploring the dataset to use, we selected the algorithm to use for mining patterns. To explain that algorithm, we first describe some pattern mining concepts.

Pattern

A pattern is a characteristic or a set of characteristics that describe a group of objects [75, 56]. That set of characteristics is usually represented by a conjunction of relational statements. For example, the pattern $[country = Canada] \wedge [hour \in [7, 11]] \wedge [conversion = true]$ refers to visitors from Canada that requested a page between 7:00 and 11:00 and made a conversion. Each pattern has correspondent support [75]. The support indicates the proportion of objects that meet the description of the pattern in a class. Consider d the number of objects in the dataset D that meet the description of the pattern p. The support of the pattern p is calculated dividing d by the total number of objects in D. The pattern used as an example could have, for example, support of 0.07 for class A and support of 0.85 for class B; which would mean that 7% of visitors in class A are from Canada, requested a page between 7:00 and 11:00 and made a conversion, and it would also mean that 85% of visitors from class B match that description [73, 76].

Contrast pattern

Contrast patterns are those whose supports differ significantly in one class with respect to the remaining classes. The example in the previous paragraph would be a contrast pattern because the percentage of objects covered by the pattern is very different between the two classes: 85% in class B and 7% in class A [73, 76, 107].

Class imbalance

We dealt with the class imbalance problem, which means that there are significantly fewer objects of one class (called minority class) with respect to another (called majority class), e.g., for contrasting visitors who made a payment against those who did not, we have a percentage of 5% and 95% respectively (See Table 3.5). When the class imbalance problem is present, contrast pattern

miners can obtain high-support patterns for the majority class. However, they only find a few low-support patterns for the minority class. There exist few contrast pattern-based classifiers that tackle this problem [73, 76, 75]. We decided to use one of those classifiers due to the imbalance problem present in our data. Next, we describe them.

Algorithms for supervised classification based on contrast patterns

Based on their **mining** strategy, the algorithms for supervised classification based on contrast patterns that have addressed the class imbalance problem could be categorised into [75]:

- **Exhaustive-search-based** (ESB): algorithms that execute a comprehensive search of values combinations for features that are significant in one class in comparison with other classes.
- **Decision tree-based** (DTB): algorithms that extract contrasts patterns using decision trees.

Considering their pattern **filtering** strategy, algorithms are based on set theory or based on quality measures for patterns. The first reduces redundancy [75].

We decided to use a DTB algorithm because ESB algorithms usually transform numerical features into nominal by creating disjoint intervals (bins) using an initial discretisation. Since that discretisation does not consider the values of other features, it could hide relations between objects. Also, contrasts patterns obtained with ESB algorithms have the symbol $=$ as the only relational operator. The initial discretisation produces information loss and reduced interpretability. Another disadvantage of some ESB algorithms is that they modify the original dataset by using resampling methods, it could create bias toward the majority class extracting patterns that are not representative of the problem. DTB algorithms for mining contrast patterns avoid those drawbacks and also reduce computational cost. Defiance when mining patterns, is the exponential number of possible patterns. Therefore, we selected an algorithm that helps us reduce the redundancy of patterns [73, 76, 75, 71, 62].

Selected algorithm: PBC4cip

Based on the mining and filtering strategy of algorithms, we selected PBC4cip. It is a contrast pattern-based classifier that tackles the class imbalance problem. Its mining strategy is decision tree-based, and its filtering method is based on set theory. Even when we did not use PBC4cip for classifying objects, it allowed us to lead with the class imbalance problem. Other advantages of using PBC4cip were avoiding the use of resampling methods and obtaining patterns that are easy to interpret [75, 73].

3.4.3 Experimental setup

We used the Weka implementation of PBC4cip. Below we mention our experimental setup:

- PBC4cip has two variants: univariate [73] and multivariate [56] decision trees. The multivariate version finds multivariate relations and has reported better classification results [56]. However, multivariate patterns may be difficult to interpret for marketing experts. We used the univariate decision tree builder because we found univariate relations easier to interpret. Below we show an example of both types of relation:

  - Univariate relation: *numberOfVisitedPages* < 17

  - Multivariate relation: *0.01 * numberOfVisitedPages + 0.01 * depth* > 0.0403

- To obtain contrast patterns for each type of conversion (class of visitors), we:

  - Used the One-vs-Rest strategy. We have 16 conversion features that correspond to the 15 conversion types listed in Table 3.2 plus the feature "conversion", which indicates if any conversion was performed. We use each conversion attribute as a binary class at a time. That means that we run PBC4cip 16 times using the dataset obtained in 3.3.2 and described in 3.4.1.

  - Contrasted visitors who performed each conversion against its complement. For example, using the feature $requestOnlineHelp$ as the binary class, we contrasted the 70% of visitors who did not request online help against the 30% of visitors who requested it.

- We used Random Forest miner and Hellinger distance as distribution evaluator. This configuration has provided good AUC results in previous research [73, 21].
- We tried different numbers of trees. We obtained contrast patterns for both classes using 50, 100 or 150 trees. As a rule of thumb, one can use 150 trees because it has provided good classification results [56].
- There are usually duplicated patterns because they are obtained from numerous decision trees generated out of the same dataset [57, 56]. To remove duplicated patterns, we used the filtering option available in the Weka implementation of PBC4cip.

With the previous process, we obtained a set of contrast patterns for each conversion. We obtained thousands of patterns. To select the best patterns we followed the process that is described next.

3.4.4 Pattern selection

We followed four steps to select the best patterns:

1. Selecting representative patterns: to select contrast patterns that are in a representative proportion of sessions, we used a minimum of 0.10 support as the threshold. This criterion drastically reduces the number of patterns. We are not interested in patters that are in less than 10% of the sessions, but a different threshold can be used. The support has been used to select contrast patterns in previous research [38, 21].
2. Removing redundant items: we eliminated pattern items that are contained in a more specific item. For example, in the pattern *sessionDuration* $\leq$ *1.5 (seconds)* $\wedge$ *sessionDuration* $\leq$ *5 (seconds)* $\wedge$ *userIsNew* = *True*, we can eliminate the item *sessionDuration* $\leq$ *5 (seconds)*, because all sessions with a duration ≤ 1.5 have also a duration ≤ 5. Then, the pattern would be simplified as *sessionDuration* $\leq$ *1.5 (seconds)* $\wedge$ *userIsNew* = *True*. This criterion has been used in previous contrast patterns research [57, 56, 74].
3. Removing specific contrast patterns: we removed contrast patterns that are more specific than other contrast patterns. A pattern P1 is more specific than a pattern P2 if P2 is contained

in P1 and P1 has at least one more item. Consider P1 = *sessionDuration* $\leq$ *1.5 (seconds)* $\wedge$ *userIsNew* = *True* $\wedge$ *madePayment* = *True* and P2 = *sessionDuration* $\leq$ *1.5 (seconds)* $\wedge$ *userIsNew* = *True*. We remove P1 because it is more specific than P2. This criterion has been used in previous contrast patterns research [57, 56, 74].

4. Selecting patterns with less number of items: the adequate number of items in a pattern is subjective but, in general, patterns with fewer items are easier to interpret. We kept only patterns with three or fewer items, which we consider are easier to transform into actionable information.

5. Interestingness: it is a subjective selection usually made by a domain expert. It allows identifying the most relevant patterns according to business objectives. For example, the pattern *initialPage* $\neq$ *.../newsletters/* $\wedge$ *requestedHelp* = *False* could be not interesting for the expert because knowing that the initial page was not ".../newsletters" leaves hundreds of possible initial pages. Nevertheless, the feature "initial page" is informative when it refers to a relevant page such as "Home", or when it comes with the relational operator "=".

The proposed criteria allow for a representative and informative set of contrast patterns. Assessing different quality metrics is not in the scope of this thesis. However, other objective and subjective quality measures can be used for selecting contrast patterns [62, 32, 70, 72]. A comparison of objective quality measures for selecting contrast patterns is in the research of Garciá-Borroto et al. [62]. There is also research, from Loyola-González et al., regarding the effect of class imbalance in quality measures [72, 70].

With the proposed criteria, we selected the best contrast patterns for each conversion. As an example, Table 3.2 shows some selected patterns using the feature *conversion* as the class. This feature indicates if at least one conversion was performed in the session or not. The support is in the form $[Class\ F \quad Class\ T]$:

- $Class\ F$ indicates the proportion of sessions in which no conversion was performed (objects with the label "False").

- $Class\ T$ indicates the proportion of sessions in which at least one conversion was performed. (objects with the label "True"), e.g. confirm payment, subscribe to the newsletter or request online help.

Table 3.2: Example of selected contrast patterns.
Class F refers to sessions in which no conversion was performed. *Class T* refers to sessions in which at least one conversion was performed. For example, the last pattern refers to visitors who visited at most ten pages. This pattern was found in 92% of visitors who did not perform any conversion and it was not found in visitors who performed at least one conversion.

N°	**Pattern**	**Support**	
		[Class F	**Class T]**
1	*country = Mexico* $\wedge$ *requestOnlineHelp = True*	[0.00	0.53]
2	*consultProductAvailability = False* $\wedge$ *requestOnlineHelp = True* $\wedge$ *numberOfVisitedPages* $>$ *25*	[0.00	0.53]
3	*userIsNew = False* $\wedge$ *numberOfVisitedPages* $>$ *37*	[0.00	0.68]
4	*sessionDuration* $>$ *1.5 (seconds)* $\wedge$ *requestOnlineHelp = True*	[0.00	0.68]
5	*knownReferrer = False* $\wedge$ *numberOfVisitedPages* $\leq$ *3*	[0.72	0.00]
6	*sessionDuration* $\leq$ *1.5 (seconds)*	[0.76	0.00]
7	*averageSecPerRequest = 0.00*	[0.81	0.00]
8	*numberOfVisitedPages* $\leq$ *10*	[0.92	0.00]

The patterns selected for each conversion could be considered the characterisation of visitors. However, we manually interpreted these patterns. In subsection 3.4.5, we present the interpretation.

3.4.5 Pattern interpretation

The last step of our methodology is the interpretation of the marketing expert in the context of the sales funnel. We manually interpreted the contrast patterns obtained for each class of visitors and validated them with a marketing expert. In this section, we present the interpretation of the patterns listed in Table 3.2. Then, we present some insights obtained from the patterns for each step of the sales funnel.

Interpretation of contrast patterns obtained using the feature *Conversion* as the class

The scope of this group of patterns is to find what distinguishes visitors who made a conversion from those who did not. We know that 58% of visitors did not perform any conversion, and 42% performed at least one conversion (see Figure 3.5).

From the visitors who did not make a conversion:

- 92% visited at most ten pages.
- 81% left each page almost immediately, on average.
- 76% were on the website at most 1.5 seconds.
- 72% came from an unknown referrer (were organic traffic) and visited at most three pages.

From the visitors who made a conversion:

- 68% were on the website for more than 1.5 seconds and requested online help.
- 68% were recurrent visitors and visited more than 37 pages.
- 53% did not consult product availability, requested help, and visited more than 25 pages.
- 53% were from México and requested online help.

Insights obtained for each step of the sales funnel

For each type of conversion, we replicated the interpretation method exemplified in the previous paragraph. In Appendix B, we summarise that interpretation. Below we present some findings grouped by sales funnel step.

In the **awareness** step, which is the first step of the sales funnel, it was expected the highest number of conversions. Nevertheless, awareness conversions have the lowest conversion rates. Even when most types of conversions (6 of 15) are targeted to this step. On average, the conversion rate of conversions in the awareness step is 0.75%. This step has a non desired conversion, which is *unsubscribed from the newsletter*. Its percentage is higher than the percentage of subscription. The web pages associated with the conversions of the awareness step require a periodical creation or renewal of content. Therefore, it could be evaluated if it is worth to maintain all of them.

In the **consideration** step, it is expected a lower number of clients than in the awareness step. Nevertheless, it was the opposite. Marketing experts could evaluate if the higher conversion rate in the consideration step is intentional (they put more resources and efforts) or if they expected awareness conversions to be more effective. The two conversions with the lowest conversion rate in the consideration step are on web pages with static content, therefore maintaining them is probably not costly, unlike the awareness conversion types. The "product A" is strongly promoted by the website. We found that visitors who visit web pages associated with the "product A" usually *logged in* and *consulted product availability*. Therefore, a more striking add related to the "product A" could be placed on the login and consult availability pages.

In **intent**, we interestingly found the highest conversion rate, even when it has only two types of conversion. That indicates that the website is effective for attracting visitors that currently need the product or service and are eager to pay for it. Marketing experts could analyse how to use the high conversion rate in the conversions of the intent step to increase the conversion rate at the next step, where the purchase occurs. In the intent step, we have the conversion *requested online help*, which has the highest conversion rate. That indicates that the online help is a relevant contact point with clients or potential clients. Since it is very close to the purchase step, the online help could be improved to increase the conversion rate at the purchase step, e.g., additional to answer the questions of visitors, proactively mention the advantages over other brands or offer discounts.

Comparing the average conversion rate of conversions in the intent and **purchase** steps, we found that the website loses about 85% of visitors with purchase intention. Marketing experts could deeply analyse the characteristics of visitors in intent and purchase steps to find out ways to narrow the conversion gap in these steps.

In the **loyalty** step, we found no conversions aimed to measure the satisfaction of clients, make clients re-purchase or make clients become promoters of the product. Marketing experts could evaluate the feasibility of adding loyalty strategies on the website, such as rewards, affiliate or referral programs, and marketing strategies like flash-offers for current clients. We found the characteristics of visitors who have already bought a product. Those characteristics could be useful for designing marketing strategies aimed to promote client loyalty.

3.5 Conclusions

In this chapter, we presented an approach for characterising visitors using conversions associated with the sales funnel as classes. We described a methodology that could be replicated using web logs from any website as input data. The result was a characterisation of different classes of visitors that is easy to interpret and could be meaningful for marketing experts.

A contribution of this research is the creation of sessions with our code instead of using individual pages. By using our code, we eliminated dependency on third-party software, and we obtained a dataset that can be transformed for different purposes. For example, besides the characterisation of visitors, we also used this dataset for describing classes of visitors with a sequence mining approach. That is explained in Chapter 4.

Another contribution is the use of conversions associated with the sales funnel as classes of visitors. We proposed this approach because *conversions* and *sales funnel* are relevant for measuring and improving the effectiveness of a website. Besides, marketing experts are familiar with both concepts. Using conversions as classes of visitors allowed us to obtain specific characteristics for visitors who performed, or did not performed, each type of conversion. By associating conversions with steps of the sales funnel, results could be used to improve the effectiveness of conversions at a given sales funnel step. That is useful because no company can serve all customers with the same level of satisfaction [82, 36].

Chapter 4

Description of the navigation behaviour of visitors

Web Log Mining is the use of data mining techniques to obtain information about the navigation behaviour of visitors ([45]). The main difficulties in Web Log Mining are the huge amount of traffic on websites and the wide variety of paths that visitors could follow ([22]). Understanding the navigation behaviour of numerous visitors can be overwhelming. Therefore, in this chapter, we aim to describe the navigation behaviour of visitors with a simplified representation. With a sequence mining approach, we captured the milestones of different classes of visitors and presented them as a graph. That reduced the amount of data that a marketing expert would have to analyse for understanding the navigation behaviour of visitors and contrast different classes of visitors.

To achieve this goal, we represented the navigation behaviour of the website visitors as the sequence of visited pages in each session. That representation allows to find out the representative navigation milestones in each class of visitors. To this aim, we used a compression algorithm that allowed us to identify sequences of pages that are common among visitors from the same class. We called those common sequences "rules". The set of rules obtained in each class of visitors describes the navigation behaviour of most of the visitors in that class.

Having identified the rules for each class of visitors, it is possible to replace the sessions pages with rules. That results in a reduced representation of sessions that allows to carry out different kinds of analysis. For example, sessions could be represented exclusively with rules or, conversely, the behaviour that is not represented by rules could be analysed further. Statistics of the sessions represented as rules would help a marketing expert to establish questions of interest. We analysed those statistics and, due to our simplification purpose, we found relevant to represent the navigation behaviour of visitors only with the most frequent rules.

For our target audience, which are marketing experts, a graph visualisation of results would be more friendly. Therefore, we summarised all the sessions of a given class in a graph. The representation of sessions with rules significantly reduced the amount of data that a marketing expert would have to analyse. The graph depiction eases to understand the navigation behaviour of visitors, even when our work was not focused on visualisation techniques.

In this chapter, we describe the Web Log Mining process that we followed. In Section 4.1, we explain the sequence mining process for identifying rules in each class of visitors. Then, in Section 4.2, we explain how we represented sessions with rules. In Section 4.3, we present the graph that describes the navigation behaviour of visitors and insights obtained from it. Finally, in Section 4.4, we present our conclusions.

4.1 Extraction of representative sequences of pages (rules)

In this section, we describe how we found out the most common sub-sequences of visited pages for different classes of visitors. In the first place, one needs to represent the navigation behaviour of website visitors. On that representation, then it is possible to find sequences of web pages that are commonly visited among visitors from a given class. Therefore, in subsection 4.1.1, we describe how we represented sessions with a sequence of symbols that contains the navigation behaviour of interest. Then, in subsection 4.1.2, we explain the compression algorithm that we used to find common sub-sequences of pages. Finally, in subsection 4.1.3 we describe how we used that compression algorithm to find out the most common sub-sequences of visited pages,

which are the milestones for each class of visitors.

4.1.1 Representation of each class of visitors as a sequence of symbols

The input data was obtained in Chapter 3. It consists of 50,820 sessions represented as the list of visited pages. There were thousands of different pages, but only a small subset was relevant for the proposed analysis. Once we identify relevant pages on sessions, we represent them as symbols to ease the sequence mining process.

Identification of relevant pages

Below we describe the steps that we followed. We are interested in the navigation behaviour that could be meaningful for marketing experts. Steps could be different depending on the analysis purposes.

1. Filtering of pages of interest: Our marketing partner prepared a list of web pages that they are interested in analysing. That list contained 298 web pages. We removed from sessions all web pages that were not pages of interest. 26% of the sessions did not visit any page of interest. We eliminated those sessions. The dataset was reduced to 37,400 sessions (74%).

2. Removing of pages that are automatically loaded: Those pages are not meaningful for marketing experts because do not represent the intentional navigation behaviour of visitors (for example, java resources necessary for the proper functioning of the site).

3. Removing the subsequent repetition of the same page: Most sessions had pages subsequently repeated n times. For example, *Home* $\rightarrow$ *Home* $\rightarrow$ *Home* $\rightarrow$ *Home* $\rightarrow$ *Login*$\rightarrow$ *Control panel* $\rightarrow$ *Control panel* $\rightarrow$ *Logout*. Our Information Technology partner indicated to us that, in most cases, it is due to the functionality of the website, not related to the navigation behaviour of visitors. For example, if the visitor fulfils a form, the same page could be automatically reloaded whenever the visitor clicks on a different field. Therefore, we reduced the subsequent occurrences of the same page to one occurrence. The previous example would be reduced to *Home* $\rightarrow$ *Login*$\rightarrow$ *Control panel* $\rightarrow$ *Logout*.

The process to keep only relevant pages entailed some information loss. That information could be useful for some traffic analytics. For example, to measure the number of pages sent to the visitor, the amount of data transmitted, or the frequency of clicks. Nevertheless, that loss does not affect the objective of describing the navigation behaviour of visitors. We only need web pages intentionally visited.

Representation of sessions as a sequence of symbols

To ease the sequence mining process, we represented each session as a sequence of symbols. Because there are 298 pages of interest, we assigned a 2-letter identifier to each of them. Then, in each session, we replaced the name of the page with its identifier. For example, the session *Home* $\rightarrow$ *Login*$\rightarrow$ *Control panel* $\rightarrow$ *Logout* became $AaAzBkBb$.

Segmentation of data in different classes of visitors

We are interested in describing the navigation behaviour of different classes of visitors and contrast them. Therefore, it is necessary to segment data. The input dataset was already labelled. We classified 100% of sessions into four disjoint classes:

- Visitors who made a payment: It corresponds to 7% of the sessions.
- Visitors who started the payment process but did not conclude it: It corresponds to 10% of the sessions.
- Visitors who made a conversion different to made payment or started payment: It corresponds to 32% of the sessions.
- Visitors who did not perform any conversion: It corresponds to 52% of the sessions.

We will refer to the previous classes of visitors as ***Made payment***, ***Started payment***, ***Other conversions*** and ***No conversion***. We obtained a dataset for each class of visitors.

With sessions represented as a sequence of symbols and segmented in different classes of visitors, it is possible to identify the representative navigation milestones in each class of visitors. To this aim, we used a compression algorithm that is described in the following subsection, 4.1.2.

4.1.2 Selection and implementation of the compression algorithm

An objective of our research was to reduce the huge amount of data that marketing experts should have to analyse for understanding the navigation behaviour of visitors. Our strategy was to find recurrent sub-sequences of visited pages. Therefore, we used a sequence mining approach. In this subsection, we explain how we selected the sequence mining algorithm, how it operates and the implementation that we used.

Selection of the sequence mining algorithm

We discarded algorithms that find the longest common sub-sequence, like MAXLEN ([19], [90]), because we are interested in all the sub-sequences that are repeated, no matter if they are long. We evaluated compression algorithms such as Sequitur ([90], [63], [35]), Repair ([66], [81]), and Bisection ([34], [90], [49]). Sequitur algorithm is the most efficient. It runs in linear time. Therefore, we selected it for finding the recurrent sub-sequences of visited pages. Below we explain this algorithm.

Sequitur algorithm

Sequitur finds repetitive sub-sequences in a sequence by identifying rules. It creates a grammar based on repeated sub-sequences. Then, each repeated sub-sequence becomes a rule in the grammar. To produce a concise representation of the sequence, two properties must be met ([26], [55]):

- $p1$ (**digram uniqueness**): there is no pair of adjacent symbols repeated in the grammar.
- $p2$ (**rule utility**): every rule appears more than once.

To clarify the operation of Sequitur, we will use the following definitions:

- **Sequence**: a string of symbols, e.g., “aghhhhbfababdchdttttyhhs”.
- **Rule**: a sub-sequence that appears twice or more in a sequence and its minimum length is 2. The rules obtained with Sequitur algorithm may be defined in terms of other rules.

- **Base rule**: a rule that does not contain other rules, e.g., rule 1 = "a b", rule 2 = "d c".
- **Nested rule**: a rule composed of base rule(s), e.g., if rule 3 = "f 1 1 2 h", it is a nested rule defined in terms of the base rules 1 and 2.
- **Expanded rule**: the result of recursively unfolding all the rules that are contained in a nested rule, e.g., the nested rule "f 1 1 2 h" is expanded as "f a b a b d c h".

We will use the sequence **aaghdfghmaadfgh** as an example to describe the operation of the Sequitur algorithm. In Table 4.1, we can see that Sequitur does not find any rule from rows 1 to 7. That is because there are no pairs of symbols that appear twice or more in the string. In row 8 the pair of symbols "gh" appears twice, so it is added to the grammar as rule 1. In row 11 the pair of symbols "aa" appears twice and it is added to the grammar as rule. In row 13 the pair of symbols "df" appears twice and it is added to the grammar as rule 3. In row 15 the rule "df" becomes a nested rule because "dfgh" is found twice, but the pair "gh" is already the rule 2, so the rule 3 changes from "df" to "df 2". All the rules added to grammar met properties $p1$ and $p2$.

Implementation of the Sequitur algorithm

We used a publicly available implementation of Sequitur ([7]). We adapted this implementation for using it with the 2-letter identifier of each web page. That was necessary because the original implementation identifies each symbol as a different element in the sequence.

In the following subsection, 4.1.3, we explain how we used the described implementation of the Sequitur algorithm for finding recurrent sub-sequences of visited pages.

4.1.3 Rule extraction

We used the Sequitur algorithm to find rules in each class of visitors. Those rules, which are recurrent sub-sequences of visited pages, are the milestones of each class of visitors. In this subsection, we explain how we extracted, analysed, and selected those rules.

Table 4.1: Operation of the Sequitur algorithm.
We use the sequence **aaghdfghmaadfgh** as an example. For each symbol in the sequence, Sequitur verifies the properties of digram uniqueness and rule utility. In each row, we show the resulting grammar and the expanded rules, as each new symbol is reviewed. In the column "Resulting grammar", 1 to n are the found rules, and 0 is the result of using those rules in the original string. Grammar 0 is not expanded in the last column because it is not a rule. But, if we expand Grammar 0 we obtain the original string.

No.	**New symbol**	**The string so far**	**Resulting grammar**	**Expanded rules**
1	a	a	0 → a	*No rules found*
2	a	aa	0 → a a	*No rules found*
3	g	aag	0 → a a g	*No rules found*
4	h	aagh	0 → a a g h	*No rules found*
5	d	aaghd	0 → a a g h d	*No rules found*
6	f	aaghdf	0 → a a g h d f	*No rules found*
7	g	aaghdfg	0 → a a g h d f g	*No rules found*
8	h	aaghdfgh	0 → a a 1 d f 1 1 → g h	 gh
9	m	aaghdfghm	0 → a a 1 d f 1 m 1 → g h	 gh
10	a	aaghdfghma	0 → a a 1 d f 1 a 1 → g h	 gh
11	a	aaghdfghmaa	0 → 1 2 d f 2 m 1 1 → a a 2 → g h	 aa gh
12	d	aaghdfghmaad	0 → 1 2 d f 2 m 1 d 1 → a a 2 → g h	 aa gh
13	f	aaghdfghmaadf	0 → 1 2 3 2 m 1 3 1 → a a 2 → g h 3 → d f	 aa gh df
14	g	aaghdfghmaadfg	0 → 1 2 3 2 m 1 3 g 1 → a a 2 → g h 3 → d f	 aa gh df
15	h	aaghdfghmaadfgh	0 → 1 2 3 m 1 3 1 → a a 2 → g h 3 → d f 2	 aa gh dfgh

Rule finding

Sequitur identifies as rules those sub-sequences that appear twice or more in a string. Nevertheless, for our analysis, it was necessary to find all sub-sequences that are common among different sessions. Some of those sub-sequences may appear only once in each session. To this aim, we concatenated sessions of each class of visitors. Below we explain the methodology that we followed for finding rules. We used the implementation described in 4.1.2.

1. Concatenate all sessions of a given class of visitors, adding a distinguishing pair of symbols between each session.
2. Apply the Sequitur algorithm.
3. Expand rules.
4. Exclude rules that include the pair of symbols mentioned in the first step.
5. Compute the frequency of each rule in sessions of the same class of visitors.

In Table 4.2, we show the percentage of sessions and the number of rules found in each class of visitors. We can see that the classes of visitors *Made payment*, *Started payment* and *Other conversions* have a much higher number of rules than the *No conversion* class of visitors, even when *No conversion* has the highest percentage of visitors. That could indicate a more homogeneous behaviour in visitors from the first three classes.

Table 4.2: Rules obtained in each class of visitors.
7% of the sessions are visitors from the class *Made payment*, where we found 764 rules. 52% of the sessions are visitors from the class *No conversion*, where we found only 92 rules.

Metric	**Class of visitors**			
	Made payment	**Started payment**	**Other conversions**	**No conversion**
Percentage of visitors	7%	10%	32%	52%
Number of rules	764	704	997	92

Rules should allow us to contrast classes. Therefore, we made an inter-class analysis to find out if the set of rules is different, or behaves differently, in each class of visitors.

Inter-class analysis

The objective of the inter-class analysis is to find out 1) if the rules are different for each class of visitors, and 2) if those rules are relevant. To this aim, we computed two metrics:

- **Percentage of rules found in sessions**: given a set of rules, it measures the percentage of those rules that are found in a group of sessions. It allows us to find out if a set of rules describes a specific class of visitors or not. A result of 100% in all classes of visitors for a given set of rules would mean that all those rules were found in the four classes of visitors. Thus, that set of rules would not describe a specific class of visitors.
- **Inverse frequency of a rule**: it measures the percentage of sessions in which a rule is found at least once. A high percentage indicates that the rule is relevant to describe the navigation behaviour of visitors.

As an example of the inter-class analysis, in Table 4.3, we show the metrics of the rules found in visitors from the class ***Made payment***. Below we summarise the interpretation of this table.

- 100% of the rules were found in *Made payment* sessions because rules were extracted from those sessions. We can see that this percentage decreases to about 50% for the classes of visitors *Started payment* and *Other conversions*. For the visitors from the class *No conversion*, it reduces to 6%. These results indicate that about 50% of the rules specifically describe the navigation behaviour of the visitors that belong to the class *Made payment*.
- The highest inverse frequency indicates that one rule was found up to 91% of sessions that belong to the class of visitors *Made payment*. This metric is lower for the other three classes of visitors. Since we use this metric to measure the rule relevance, we can say that this set of rules is more relevant for the visitors that belong to the class *Made payment*.

Table 4.3: Inter-class comparison of the rules found in "Made payment" class.

Metric	**Class of visitors**			
	Made payment	**Started payment**	**Other conversions**	**No conversion**
Percentage of "made payment" rules found in sessions.	**100%**	57%	46%	6%
Highest inverse frequency of a "made payment" rule in a session.	**91%**	68%	33%	2%

We made the inter-class analysis for the other three classes of visitors. Both metrics are higher when the set of rules and the sessions belong to the same class of visitors. Nevertheless, it was remarkable that the highest inverse frequency of *No conversion* rules was only 14% in the sessions of the same class. It indicates that the behaviour of the visitors that belong to the class *No conversion* is less homogeneous.

The inter-class analysis confirmed that there are relevant and specific rules for each class of visitors. The next step was to select the best rules for describing the navigation behaviour of visitors.

Rule selection

A selection of rules is necessary because the rules obtained until here include base rules and nested rules. That was redundant because base rules are contained in nested rules. There could also be rules with too low inverse frequency (e.g. rules that are found in just one session). Those rules are not representative. Therefore, we applied two criteria for selecting rules:

1. Select only nested rules. It eliminates redundancy.

2. Select rules with inverse frequency $\geq$5%. A rule that describes less than 5% of the sessions does not generalise the navigation behaviour of visitors. It is not representative and, thus, it is not useful for the objectives of our research.

In Table 4.4, we show the number of rules obtained after applying these selection criteria. The inverse frequency of all the nested rules extracted from the class of visitors *No conversion* was

$< 5\%$. We concluded that the navigation behaviour of this class of visitors is non-homogeneous. Thus, it could not be simplified using a small set of rules. In the next steps of the process, we only used the classes of visitors *Made Payment*, *Started payment* and *Other conversions*. From now on, when we use the term "rule(s)" we refer to the set of rules presented in Table 4.4.

Table 4.4: Rules selected for each class of visitors.
These are the nested rules with inverse frequency $\geq$5%.

	Made payment	**Started payment**	**Other conversions**	**No conversion**
Number of rules	9	5	4	0

We assigned a name to each rule. In Table 4.5, we list that name and the number of pages that form each rule. We also mention the class of visitors in which the rule was found. The rules listed in Table 4.5 are the navigation milestones for each class of visitors. Now, those rules can be used to simplify the representation of sessions. That process is explained in the next Section, 4.2.

4.2 Representation of sessions with rules

At this point, we already identify the rules for each class of visitors. These rules can be used for representing sessions. This reduced representation allows to carry out different kinds of analysis. Therefore, in Subsection 4.2.1, we explain how we used rules for representing sessions. Then, in Subsection 4.2.2, we provide statistics of the sessions represented with rules. These statistics provide information that would help Marketing or Information Technology experts to establish questions of interest. For our particular case of study, it was relevant to represent the navigation behaviour of visitors only with the most frequent rules.

4.2.1 Use of rules for representing sessions

We used the rules identified in Section 4.1 to represent sessions. In each session or group of sessions, we use only the rules that belong to the same class of visitors, according to Table 4.5. For example, a session that belongs to the class of visitors "Made payment" is represented only

Table 4.5: Name of the rules found in each class of visitors.
The rule length indicates the number of pages that form each rule. The class of visitors is indicated with a letter: M = *Made payment*, S = *Started payment*, O = *Other conversions*. A grey cell indicates that the rule was found in that class of visitors.

Rule name	**Rule length**	**M**	**S**	**O**
Go to control panel	4	■	■	■
Pay via control panel	7	■		
Pay and modify product	5	■		
Consult availability and pay	8	■		
Modify product and pay	8	■		
Start payment	4	■		
Pay for a service	8	■		
Make invoice	4	■		
Login and modify product information	4	■		■
Modify product information	3		■	■
Consult payment details	3		■	
Make payments query	3		■	
Modify product information and start payment	3		■	
Consult availability	3			■

with the nine rules found in the sessions of the same class of visitors. In Figure 4.1 we show an example of three representations of a session: 1) the original session, 2) the session we get by replacing in the original session frequently occurring subsequences with rules (we call it a shrinked session), and 3) the session we get by stripping off any symbol but a rule in a shrinked session (we call it a stripped session).

The rules represent the behaviour that is common among visitors from the same class. Conversely, pages that do not form rules represent uncommon behaviour. Having the sessions of each class represented with rules, a variety of analysis could be performed. To determine which analysis is worth to do, we obtained statistics about sessions represented with rules. Those statistics are presented in the next subsection, 4.2.2.

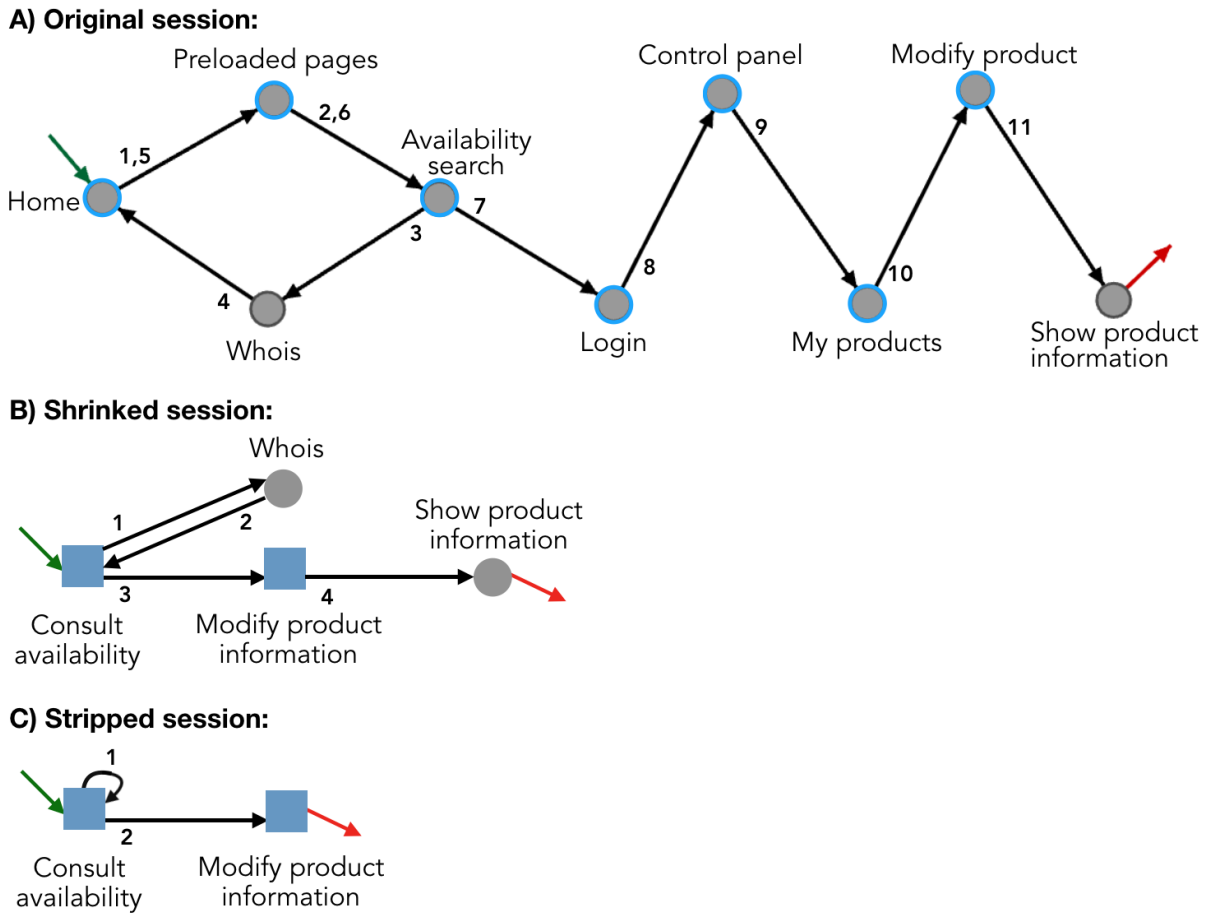

Figure 4.1: Example of a session represented with rules.
Figure A) shows the original session. The green (respectively red) arrow indicates the entry (respectively exit) page. Circles with a blue border are pages that are part of a rule. This session belongs to visitors from the class "Other conversions". Therefore, we only use the rules identified in that class of visitors. By representing the session with rules, we obtain figure B). Figure C) is the result of removing all pages that do not form a rule.

4.2.2 Statistics of different representations of sessions with rules

In this subsection, we present statistics of the sessions represented with rules. These statistics would help a Marketing or Information Technology expert to determine questions of interest. We computed statistics on the three representations exemplified in Figure 4.1: original session, shrinked session, and stripped session. We computed the length of each representation and the reduction rate with respect to the length of the original session. Using the example in Figure 4.1, the length of the original session is 12, the length of the shrinked session is 5, and the length of

the stripped session is 3.

- The reduction rate of the **shrinked session** is equal to *1-(length of the shrinked session / length of the original session)*. That is $1 - (5/12) = 0.58$. The length of the session is reduced by 0.58 (58%) when it is expressed with rules and pages that do not form a rule.
- The reduction rate of the **stripped session** is equal to *1-(length of the stripped session / length of the original session)*. That is $1 - (3/12) = 0.75$. The length of the session is reduced by 0.75 (75%) when it is expressed only with rules.

For each class of visitors, we computed the length of the three representations and the reduction rate. As an example, in Table 4.6, we show the results for the class of visitors "Made payment". We can see that the average reduction rate is 0.54 in shrinked sessions. For stripped sessions, the average reduction rate is 0.95. To better understand how the reduction rate behaves, we obtained a histogram of the reduction rate. In Figure 4.2, we show the histogram for shrinked sessions. In Figure 4.3, we show the histogram for stripped sessions.

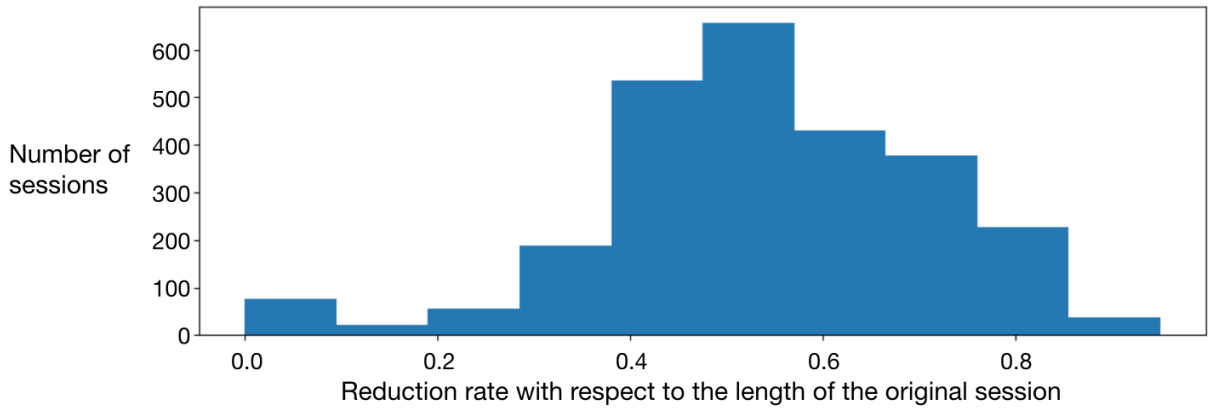

Figure 4.2: Histogram of the reduction rate in shrinked sessions.
These sessions correspond to visitors from the class "Made payment". The reduction rate is calculated with respect to the length of the original session, and it varies from zero to almost 1.

Table 4.6: Statistics of the length of sessions represented as rules.
Metrics in rows 1 to 3 refer to the length of sessions in each representation. Metrics in rows 4 and 5 refer to the reduction rate with respect to the original session. In the last row, we indicate the percentage of sessions that we used in calculations. 30% of sessions do not include any rule. Thus, in the last column, the percentage is reduced to 70%.

Metrics for visitors from the class "Made Payment"	**Original session**	**Shrinked session**	**Stripped session**
Average session length	39.23	18.26	1.67
Maximum session length	572	181	24
Standard deviation of session length	30.81	16.46	1
Average reduction rate from original session	NA	0.54	0.95
Standard deviation of reduction rate from original session	NA	0.17	0.04
Percentage of sessions considered	100%	100%	70%

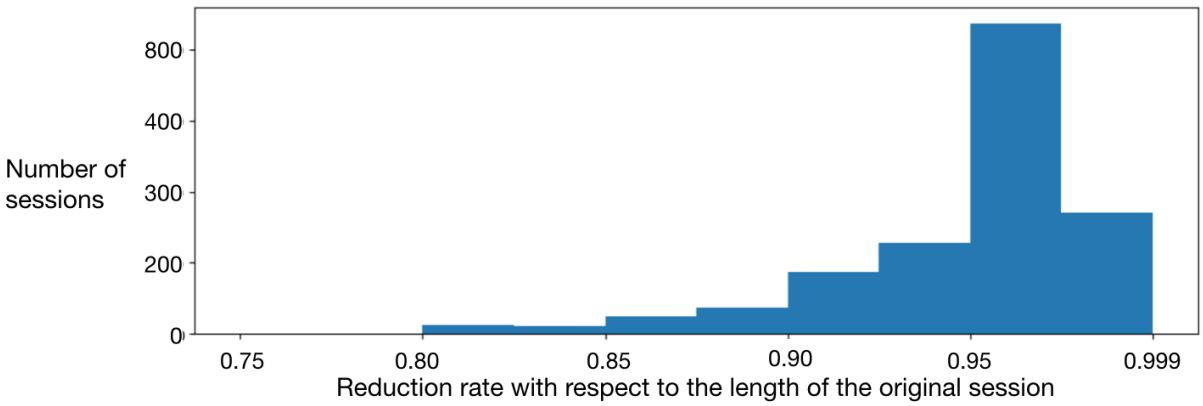

Figure 4.3: Histogram of the reduction rate in stripped sessions.
These sessions correspond to visitors from the class "Made payment". The reduction rate is calculated with respect to the length of the original session, and it varies from 0.8 to almost 1.

4.2.3 Selection of the session representation to visualise

Previous statistics would help Marketing or Information Technology experts to determine questions of interest. For example: what is the common navigation behaviour in each class of visitors?, what is different in the navigation behaviour of each class of visitors?, what navigation behaviour

is common among all classes of visitors?, what are the relevant entry and exit milestones (rules) in each class of visitors? etc. Different session representation is useful for each question.

In our case, the objective is to capture the milestones of different classes of visitors in a reduced representation of their navigation behaviour. Therefore, for further analysis, we use the shrinked sessions. It allows us to reduce the amount of data to analyse and contrast different classes of visitors. The elimination of pages that do not form rules does not affect our objective. On the contrary, including them would introduce information about the individual navigation behaviour of visitors. Nevertheless, the analysis of pages that do not form rules could be relevant for other purposes. For example, to find out what distinguishes visitors from the same class.

Using shrinked sessions, we created a graph visualisation that allows us to summarise the navigation behaviour of each class of visitors. That visualisation is presented in Section 4.3.

4.3 Visualisation of sessions represented with rules

Shrinked sessions contain the milestones of the navigation behaviour of visitors. In this section, we present those shrinked sessions in a graph visualisation aimed at our target audience, who is marketing experts. Our work was not focused on visualisation techniques, but the use of a graph is user friendly. It also allows to analyse and compare different rules or different classes of visitors. In Section 4.3.1, we explain how to build the graph. In 4.3.2, we describe the visualisation of a whole class of visitors. Then, in 4.3.3, we exemplify the analysis of a single rule. Finally, in 4.3.4, we contrast different classes of visitors.

4.3.1 Graph creation

Some calculations are necessary to build a graph that describes the navigation behaviour of visitors. In this subsection, we describe the concepts that we used to make those calculations. We also use an example to clarify them and we show the resultant graph.

Definitions

Consider i and j rules in the class of visitors A:

- **Entry rate of the rule** i: it is the number of sessions that start in the rule i divided by the total number of sessions in the class A. It is denoted $r(e_i)$.
- **Exit rate of the rule** i: it is the number of sessions that end in the rule i divided by the total number of sessions in the class A. It is denoted $r(x_i)$.
- **Frequency of the edge** $e\{i, j\}$: it is the flow of visits from the rule i to the rule j. It is equal to the number of occurrences of the edge $e\{i, j\}$. It is denoted f_{ij}.
- **Out-degree frequency of the rule** i: it is the flow of visits that goes out from the rule i. It is the sum of edge frequencies in which the source rule is i plus $r(x_i)$. It is denoted O_i.
- **Weight of the edge** $e\{i, j\}$: it is the frequency of the edge $e\{i, j\}$ divided by the out-degree frequency of the source node i, that is $\frac{f_{ij}}{O_i}$. It is denoted w_{ij}.

Calculations example

We will use an example to clarify the previous definitions. Consider the rules a and b found in the class of visitors "A". This class has 10 sessions. Also, consider the following information:

- Entry rule: 7 sessions started in rule a and 3 sessions started in rule b.
- Exit rule: 2 sessions ended in rule a and 8 sessions ended in rule b.
- Edge frequency: $f_{ab} = 5$, $f_{aa} = 3$ and $f_{ba} = 6$.

To construct the graph, it is necessary to calculate the entry rates, the exit rates, and weights.

- Entry rate: $r(e_a) = 7/10 = 0.7$ and $r(e_b) = 3/10 = 0.3$.
- Exit rate: $r(x_a) = 2/10 = 0.2$ and $r(x_b) = 8/10 = 0.8$.
- The calculation of weights is shown in Table 4.7.

Graph example

In Figure 4.4, we show the graph obtained in our example.

Table 4.7: Example of weight calculation.
The out-degree frequency O_i is the sum of frequencies f_{ij} of edges with the same source rule. Thus, $O_a = 5 + 3 + 2 = 10$ and $O_b = 6 + 8 = 14$.

Source rule i	**Target rule** j	f_{ij}	O_i	$w_{ij} = \frac{f_{ij}}{O_i}$
a	b	5	10	5/10 = 0.5
a	a	3	10	3/10 = 0.3
a	Exit	2	10	2/10 = 0.2
b	a	6	14	6/14 = 0.43
b	Exit	8	14	8/14 = 0.57

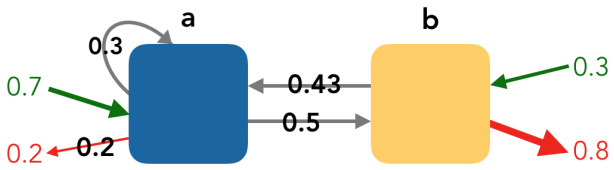

Figure 4.4: Graph example.
Yellow nodes are the rules where conversion occurs. The arrow thickness corresponds to the edge weight. The green (respectively red) arrows indicate the entry (respectively exit) rate. The values in the middle of the arrows indicate the weight of the edge (w_{ij}). If there were edges with a weight < 0.05, they would be in a lighter grey, and their weight would not be shown.

4.3.2 Visualisation of a whole class of visitors

Following the process described in 4.3.1, we created the graph for each class of visitors. In Figure 4.5, we show the graph of visitors that belong to the class "Made payment". The marketing expert could determine if the observed behaviour is expected or if there is suspicious or interesting behaviour that is worth to investigate. The interpretation of the graph depends on the business questions in which the marketing expert is interested. Below we present our interpretation.

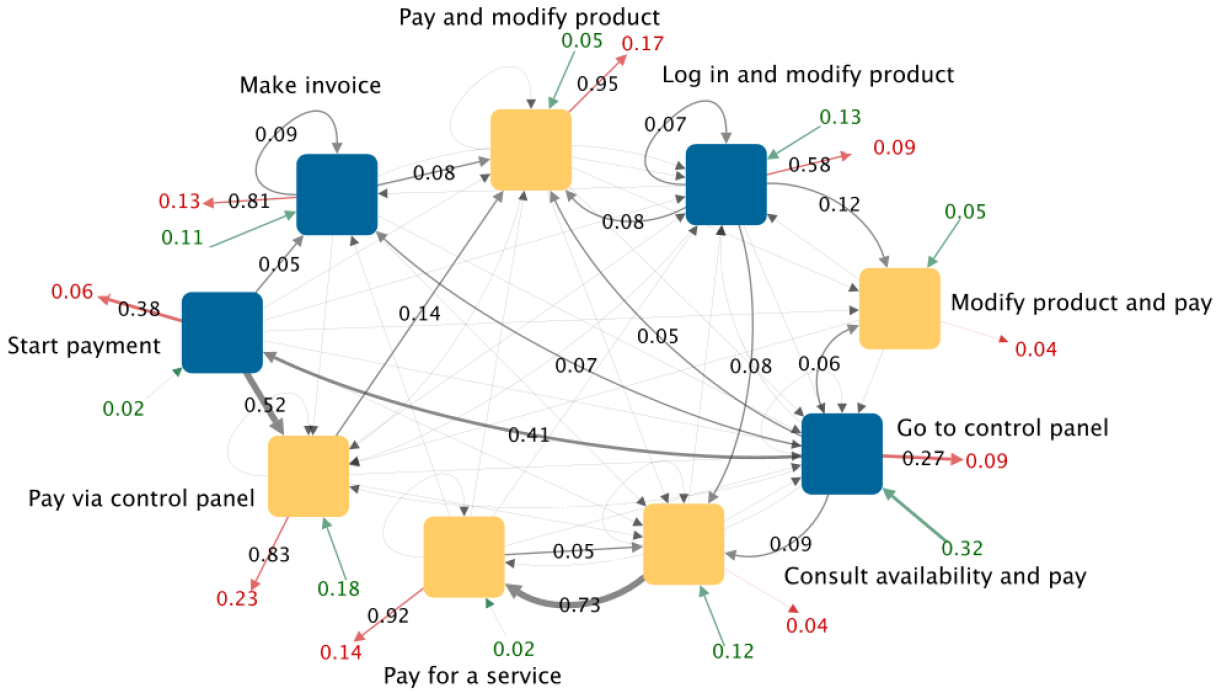

Figure 4.5: Graph of shrinked sessions for visitors from the class "Made payment". Yellow nodes are the rules where the payment is confirmed.

Relevant entry and exit rules

In Figure 4.5, we can see that the rule "Go to control panel" has the highest entry rate (0.32). 32% of the sessions have this rule as the entry point. The rule "Pay via control panel" has the highest exit rate (0.17). 17% of the sessions have this rule as the exit point.

Based on the weight of the edges, there are three rules with a weight > 0.90 on their red arrow. Those rules are "Pay and modify product", "Pay for a service", and "Modify product and pay". That means that almost all visitors who follow those rules leave the website after that. Contrarily, only 25% of visitors who follow the rule "Consult availability and pay" leave the website.

Most frequent path

If we follow the path of the highest entry rate and highest weights, we can see that 32% of visitors entry by the rule "Go to control panel". From there, 41% of visitors follow the rule "Start payment". Then, 52% of visitors follow the rule "Pay via control panel". In this rule, the payment

is confirmed. After that, 83% of visitors leave the website. The marketing expert could evaluate if this path was expected. For example, is that sequence of 15 pages adequate? Could it be shorter? Was it expected that visitors leave the website immediately after the payment?

Rules in which conversion occurs

From all the rules in which the payment occurs, most visitors leave the website. The marketing expert could determine feasible strategies to retain visitors after a purchase. For example, a flash discount on the purchase of additional service. There is an exception in the rule "Consult availability and pay". From this rule, 73% of visitors continue with the rule "Pay for a service". Those visitors made two payments because in both rules the payment is confirmed.

Besides observing the "big picture" in the graph, it is also possible to analyse specific rules in more detail. In the next section, 4.3.3, we exemplify it.

4.3.3 Analysis of specific rules

From the rules in which the payment does not occur, the rule "Make invoice" has the highest exit rate. Therefore, we will analyse this rule further. Visitors who follow this rule mainly come from the rules "Start payment" and "Go to control payment". In those rules, the payment has not been confirmed. Additional analysis from marketing expert is needed to determine the reasons for the described behaviour. For example, is the making of invoice clear? Is it a long process? Does it have an annoying bug? Does it require redundant information? Is it used by clients or competing companies to inquire about the prices of products or services?

After knowing the reasons for losing visitors in the rule "Make invoice", marketing experts could design strategies for retaining them. For example, proactive online help. If the marketing team has not enough information to determine why visitors are leaving after this rule, different actions could be implemented. For example, a pop-up window to rate the process to make the invoice. Even users who confirm the payment may provide useful information about this process.

Besides reviewing rules in detail, the graph representation also allows us to compare different classes of visitors. That is exemplified in the next section, 4.3.4.

4.3.4 Contrasting different classes of visitors

The comparison of different classes of visitors depends on the behaviour in which the marketing expert is interested. Below, we present how we contrasted the classes of visitors "Made payment" (shown in Figure 4.5) and "Started payment" (shown in Figure 4.6).

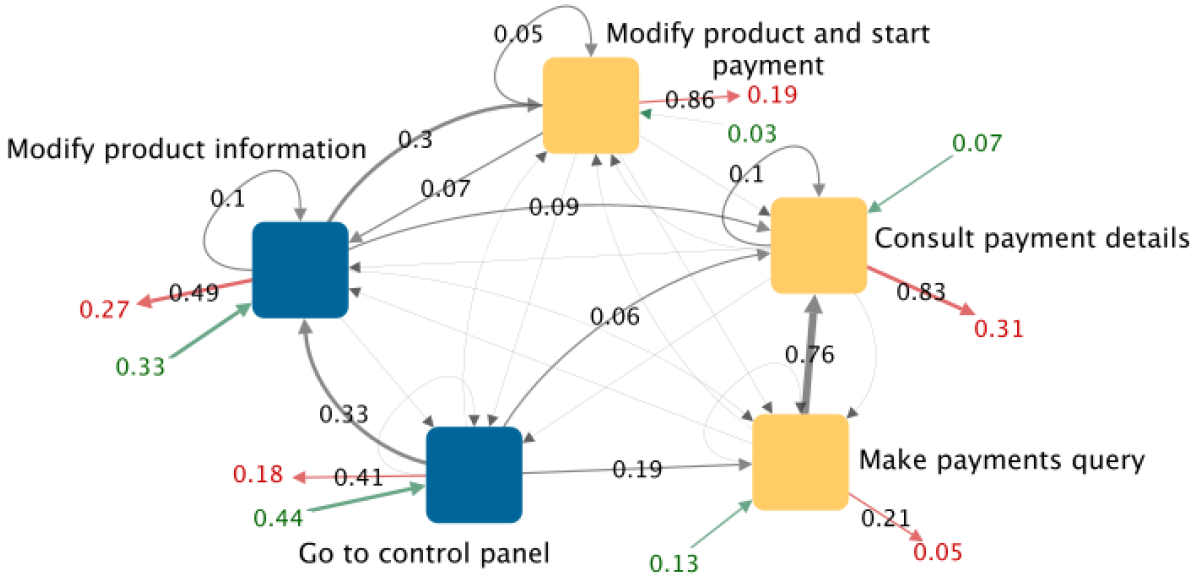

Figure 4.6: Graph of shrinked sessions for visitors from the class "Started payment". Yellow nodes are the rules where the payment is started.

Contrasting the exit rule

Most visitors from the class "Made payment" leave the website after following the rule "Pay via control panel". It is a rule in which the payment is confirmed. Most visitors from the class "Started payment" leave the website after following the rule "Consult payment details". It is a rule in which visitors start the payment. It indicates that most visitors who start a payment but do not confirm it leave the website immediately after consulting the payment details instead of navigating further or requesting online help.

In both classes, the highest exit rate is in the rules in which a conversion is performed, even though the conversion is different in each class of visitors.

Contrasting a common rule

In both classes of visitors, the rule "Go to control panel" has the highest entry rate. Nevertheless, the exit rate is double in visitors from the class "Started payment". After following the rule "Go to control panel", most visitors from the class "Made payment" start the payment process. While most visitors from the class "Started payment" modify the product information. It could be useful to encourage the purchase in the pages where the product information is modified.

Contrasting the most frequent path

The path with highest entry rate and weights, in visitors from the class "Made payment", is "Go to control panel" (32%) $\rightarrow$ "Start payment" (41%) $\rightarrow$ "Pay via control panel" (52%) $\rightarrow$ Exit (83%). In visitors from the class "Started payment", the path with highest entry rate and weights is "Go to control panel" (44%) $\rightarrow$ "Modify product information" (41%) $\rightarrow$ Exit (49%). It confirms the relevance of the rule "Modify product information" as an exit point.

4.4 Conclusions

In this section, we presented the graph of shrinked sessions for different classes of visitors. This visualisation reduces the amount of data that marketing experts would have to analyse for understanding the navigation behaviour of visitors. It also eases to contrast different classes of visitors.

Our main contributions were: 1) the use of classes of visitors associated with conversions of the sales funnel, and 2) a simplified description of the navigation behaviour of visitors by using rules. The use of classes of visitors based on the sales funnel conversions, allow us to obtain results that are easy to interpret and could be meaningful for marketing experts. The use of rules allows us to reduce the amount of data, facilitating the analysis of results.

We do not intend to replace web analytics software. It is essential to measure the traffic of a website and follow up marketing campaigns. Nevertheless, the standard configuration and reporting options of commercial software may difficult to extract high-level knowledge about the navigation behaviour of different classes of visitors. Especially due to the high amount of traffic that a website receives.

Chapter 5

Conclusions and future work

It is becoming easier and easier to create campaigns to attract visitors to a website. However, not all visitors provide an immediate business benefit. Some visitors are ready to buy a certain product. Others need more information to make a decision. There are also visitors who will never purchase the product. Attracting more visitors might be easy, but attracting the right visitors requires in-depth knowledge of them. As part of this knowledge, in this research, we proposed to find out the characteristics of visitors who perform each conversion and describe their navigation behaviour. The above are the two main contributions of this research. They are also the basis for further work, which we describe in this chapter.

To find out the distinguishing characteristics of the visitors who performed each conversion we proposed a pattern mining approach. We considered each type of conversion as a class. Then, we obtained a set of contrast patterns for each type of conversion. This approach allowed us to obtain results that are easy to interpret and could be meaningful for marketing experts. The above could be a starting point for the future development of tools that allow marketing experts to replicate the process we followed.

To describe the navigation behaviour of visitors, we proposed a clickstream analysis. It is based on identifying actions that are repeated by users of the same class. Considering an action as a sequence of visited pages. To ease the interpretation of results to marketing experts, we created a graph representation. This visualisation reduces the amount of data that marketing experts would

have to analyse for understanding the navigation behaviour of visitors. Our proposal is a starting point to further simplify the analysis of the navigation behaviour of visitors or the extraction of knowledge for a marketing audience.

This chapter is composed of two parts. In Section 5.1, we present our conclusions and future work related to the characterisation of visitors. In Section 5.2, we describe the conclusions and future work about the description of the navigation behaviour of visitors.

5.1 Characterisation of visitors using conversions as classes

The use of conversions as classes of visitors allowed us to obtain specific characteristics for visitors who performed, or did not performed, each type of conversion. In subsection 5.1.1, we present our conclusions from the process we followed. Then, in subsection 5.1.2, we present the future work that could be developed from our results.

5.1.1 Conclusions

We obtained a set of contrast patterns for each step of the sales funnel and each type of conversion. It is more informative than having a single aggregated metric. For example, we found that 42% of the visitors performed at least one conversion. Nevertheless, the percentage of visitors who performed each conversion ranges from 0% to 30%. Aggregated metrics may hide relevant information. Knowing the conversion rate for each type of conversion and the characteristics of visitors who performed them could help to identify strengths and weaknesses in each conversion. That is relevant because the methods to motivate a client to move from one point to another are different at each step of the sales funnel. As an example, based on the conversion rate, we found that the website is effective for getting visitors with purchase intention, but most of them are lost before closing the purchase.

Four main aspects allowed us to obtain contrast patterns for the visitors who performed or did not perform each conversion:

1. The identification of descriptive feature space for the sessions. It allowed us to obtain patterns that are easy to interpret.

2. The selection of the contrast pattern algorithm. We selected PBC4cip, a contrast pattern-based classifier that tackles the problem of imbalanced classes [73]. It allowed us to obtain patterns in the majority and the minority class. It also avoided the use of resampling methods and an initial discretisation of numerical features.

3. The use of conversions as classes of visitors associated with the steps of the sales funnel. It allowed us to characterise the visitors who performed each type of conversion instead of obtaining a generalised characterisation.

4. The characterisation of sessions instead of individual pages. It allowed us to obtain features that explain how visitors navigate the website (e.g. the duration of the session and the number of visited pages).

5.1.2 Future work

The PBC4cip algorithm provides contrast patterns that are easy to interpret and ordered based on their support. Nevertheless, it requires additional manual filtering based on the relevance or informativeness of patterns. This task could be automated, for example, by using a measure of the feature relevance. A disadvantage of that approach is that those criteria may be subjective, or specific for one industry or even for each website. A method could be developed for automating the filtering of patterns without needing specific business rules or domain knowledge.

There are objective quality measures that might work better than support [62, 72, 70]. In-depth research would be useful to compare different objective metrics in the context of web log mining and using data from different websites. There is also a wide variety of subjective measures that could be used to select contrast patterns. It would be valuable to test a subjective metric oriented to measure how well a pattern describes visitors at each stage of the sales funnel.

Regardless of the quality measures used, it would be important to incorporate a formal method

to validate the interpretability of patterns. There exists proven methods that could be used. For example, the Delphi method, which is used to reach experts consensus. This method is robust because, besides requiring the opinion of three experts, it also considers their grade of expertise [40, 27]. We were limited to one marketing expert, but the consensus with more experts who validate the results would be beneficial.

The offline processing that we used allows analysing historical data, which usually is not possible in web analytics software. Changes in tracking configuration usually apply for analysing future traffic. Nevertheless, the possibility of performing online analysis would also be useful (what are the characteristics of visitors who are on the website at a given moment). That would allow marketing experts to take timely decisions.

The development of a software tool where users can replicate and personalise the machine learning process that we followed could be very useful. For example, they might be interested in analysing a specific sales funnel step or characterising visitors based on a particular feature. Thus, the software should provide options to select which classes to contrast and which features to use. Bridge the gap between Machine Learning and domain experts.

The sales funnel, which is different for each company, usually includes information about all the channels that the company uses in each step. The website is only one of such channels, but it is part of a broader strategy. For example, the website could be aimed to attract new clients, and the closing of sells is performed by other channels such as e-mail or telephone. Therefore, there is also an opportunity for Machine Learning approaches aimed to analyse data from different channels of the sales funnel.

5.2 Description of the navigation behaviour of visitors

To describe the navigation behaviour of visitors, we identified actions that are repeated among visitors from the same class. It allowed us to understand how those visitors navigate. Our approach also enabled us to provide this knowledge to the marketing expert in a single graph. In

subsection 5.2.1 we present our conclusions from the process we followed. In subsection 5.2.2 we present the future work that could be developed from our results.

5.2.1 Conclusions

There are three main advantages of our methodology over other existing solutions. The first advantage is that it reduces the amount of data to analyse for understanding the navigation behaviour of visitors. The second advantage is that it extracts the real navigation behaviour of visitors. The third advantage is the use of web logs as entry data. Below we explain them in more detail.

The increasing amount of data generated on websites makes it difficult to find relevant knowledge. With our method, we replace tens of pages with a single graph. Besides summarizing the navigation behaviour of a whole class of visitors, our approach also allows us to compare different classes of visitors. This knowledge could be used to improve the effectiveness of marketing campaigns or website design. For example, an action (which is composed of a sequence of pages) could be especially successful to attract new visitors, but unsuccessful to make clients purchase. Marketing experts could design strategies for visitors to leap from the interest stage to purchase (e.g. add proactive online help, provide more information about the benefits of the product, or a retargeting campaign for the visitors who performed visited that sequence of pages). With our methodology, we extract the real navigation behaviour of visitors. It distinguishes our work from other solutions. Previous approaches group the web pages in tasks identified by the business expert. Therefore, they reflect the expected behaviour, not the real paths followed by visitors. Our approach, on the contrary, obtains the common sub-sequences of visited pages that visitors follow. No matter if those sub-sequences were expected or not. This approach allowed us to find useful information, for example, that the making of the invoice is a relevant exit point. It also enabled us to contrast relevant entry and exit rules for different classes of visitors.

The use of web logs as entry data allows performing a retrospective analysis. This is not possible in commercial software, were the configuration of conversions and market segments usually applies only for future traffic of the website. The use of web logs also allows preparing data according to different objectives. For example, we could compare different periods of a given class

of visitors or different website versions.

5.2.2 Future work

There is a latent need for creative ways to help business experts evaluate the performance of a website. For example, it could be tested the effectiveness of our approach for improving metrics measurement, website design and paid marketing effectiveness. It could be useful software aimed at marketing experts for autonomously replicating and personalizing the process that we followed. For example, the company could identify the 5 most relevant conversions and use them for describing the navigation behaviour of visitors. On the contrary, the company may find it useful to associate each page of interest with a conversion.

It would also be valuable to extract patterns from the high-level visualisations that we obtained. For example, Acosta-Mendoza et al. propose a frequent approximate subgraph mining approach [68], which we could incorporate as the last step of our methodology.

Appendix A

Feature description

In Chapter 3 we proposed the feature space for describing sessions. Below we list the name and description of each proposed feature. Features are grouped in seven categories: session timing, geographic, website navigation, type of visitor, hardware and software, size and format of visited pages, and request and server response.

1. Session timing

Day of the week: It refers to Monday, Tuesday, Wednesday, etc. Unlike the original date format, this can be used to obtain patterns.

Hour: The hour without considering minutes, that is a number from 1 to 24. This form is easier to use it to find patterns.

Session duration (seconds): The seconds that a user spent on the website, including all the requests that are part of the session.

Average time per page (seconds): The average number of seconds between each request, that is: session duration / number of requests.

2. Geographic

City: The city in which the session occurred.

Country: The country in which the session occurred.

Subdivision: The subdivision in which the session occurred. It is more specific than the city.

Organization: The organization in which the session occurred, when it exists.

3. Website navigation

Depth of visited pages: The highest number of / symbols in the URL of the visited pages. Each / symbol indicates a web site level, considering a hierarchical order in web pages. E.g. the URL home/courses/course1 has a depth = 2.

Number of requests per session: The number requests that integrate the session, each request corresponds to a requested web page.

Known referrer: A boolean attribute that indicates if the visitor comes from a known referrer.

Referrer URL: The URL from where the visitor arrived on the website, if it exists. It is commonly used to track marketing campaigns and measure its effectiveness.

Entry page The first page visited in the session.

Exit page The last page visited in the session.

4. Type of visitor

New visitor: It indicates if a visitor is new, based on its ”fingerprint” which consists of the IP address + agent device + agent operating system.

Known bot: It indicates if the requests of the session came from a known bot.

5. Hardware and software

Web browser: It is the web browser used to navigate on the website, e.g. Firefox, Safari, Chrome, etc., without the version number.

Operating system: It is the operating system used to navigate on the website, e.g. Windows, Mac OS, Ubuntu, Linux, etc., without the version number.

Device: It is the device brand used to navigate on the website, e.g. iPhone, Samsung, Huawei, etc., without model and version number.

6. Size and format of visited pages

Bytes downloaded per session: The volume of data transferred to the web client (in bytes).

Requested robots file: It indicates if any request of the session includes the robots.txt file.

Percentage of pages that are image files: The number of image files (e.g. png, jpg, do) divided into the total number of requests of the session.

Percentage of java server pages: The number of java server pages requested in one session divided into the total number of requests of the session.

Percentage of pages that are HTML files: The number of HTML files divided into the total number of requests of the session.

Percentage of pages that are pdf files: The number of pdf files divided into the total number of requests of the session.

7. Request and server response

Percentage of requests that use the GET method: The number of requests that asked for retrieving data without modifying it, divided into the total number of request of the session.

Percentage of requests that use the HEAD method: The number of requests that asked for transferring the header section only, divided into the total number of request of the session.

Percentage of requests that use the POST method: The number of requests that sent data to the server using HTML forms (e.g. file uploading), divided into the total number of request of the session.

Percentage of informational status (1xx): The number of requests with informational status (1xx) in a single session divided into the total number of requests. E.g. a status that indicates that the request was received and understood.

Percentage of success status (2xx): The number of requests with success status (2xx) in a single session divided into the total number of requests. E.g. the requested action was accepted.

Percentage of redirection status (3xx): The number of requests with redirection status (3xx) in a single session divided into the total number of requests. E.g. when a web page has migrated and is the old URL is redirected to the new one.

Percentage of client error status (4xx): The number of requests with client error status (4xx) in a single session divided into the total number of requests. E.g. when the client did not produce a request within the established wait time.

Percentage of server error status (5xx): The number of requests with server error status (5xx) in a single session divided into the total number of requests.E.g. when the server is overloaded or down for maintenance.

Appendix B

Contrast patterns interpretation

In this Appendix we summarize the interpretation of the contrast patterns that we obtained for each stage of the sales funnel. They were obtained following the process described in Chapter 3.

1. Awareness

- From the visitors who did not search for a newsletter, 59% did not perform any other type of conversion.
- From the visitors who searched for a newsletter, 60% came from an unknown referrer, requested help and consulted product information.
- From the visitors who did not visit the news section, 58% did not request help, 42% did not start a payment process, and 30% were new visitors.
- From the visitors who did not visit the glossary, 58% did not perform any other conversion, 48% were less than one second on each visited page and 46% visited less than 3 pages.
- From the visitors who visited the tutorials, 32% were from México and did not perform any other conversion.

2. Consideration

- From the visitors who did not consult product availability, 65% did not perform any other conversion, 53% were in the webpage less than 5 seconds, 43% visited less than 14 pages, 35% were from México, and 33% were new visitors.

- From the visitors who consulted product availability, 32% were recurrent visitors, and 14% were recurrent visitors from México that came from an unknown referrer, did not search new product information and did not start a payment.
- From the visitors who did not visit services description, 59% did not perform any other conversion, 52% were in the website less than 42 seconds, 50% visited less than 5 pages, and 30% were new visitors.
- From the visitors who did not consult information about the product A, 92% did not consult the availability of product, did not log in and did not make a payment, 76% did not request help, 66% did not perform any other conversion and 64% were recurrent visitors.
- From the visitors who consulted information about the product A, 84% logged in and consulted availability of product.
- From the visitors who did not consult the payment rates, 59% did not perform any other conversion, 38% were in the website less than 1 second, 29% were recurrent visitors that did not consult services information, and 25% were new visitors that visited less than 3 pages.

3. Intent

- From the visitors who did not request help, 69% did not log in and came from an unknown referrer, and 83% did not make any other conversion from which 42% were new visitors and 41% were recurrent visitors.
- From the visitors who requested help, 12% were recurrent visitors that visited more than 37 pages, came from an unknown referrer, did not log in, and did not consulted new product info nor product availability.
- From the visitors who did not start the payment process, 65% did not make any other conversion and visited less than 41 pages, 55% were on the website less than 12 seconds and were less than 1 second on each page, 41% did not request help and were on the website less than 2 seconds, 35% were from México, 32% were recurrent visitors, and 31% were new visitors that stayed on the website less than 6 seconds.
- From the visitors who started the payment process, 24% made a payment, were on the website more than 11 seconds, visited more than 61 pages and used Chrome browser.

4. Purchase

- From the visitors who did not create their user account, 61% requested less than 62 pages and were on the website less than 14 seconds, 58% did not perform any other conversion, 55% were less than one second on each visited page and did not visited on Thursday, 32% did not request help and were from México, and 32% were new visitors.
- From the visitors who created their user account, 15% made payment, consulted availability of product, did not visited on Monday, and were on the webpage more than 27 minutes.
- From the visitors who did not make a payment, 61% did not perform any other conversion, 61% visited less than 69 pages, 54% were in the website less than 86 seconds, 50% were less than one second on each visited page, 33% were from México and 31% were new visitors.

5. Loyalty

- From the visitors who did not log in, 70% did not perform any other conversion, 69% visited less than 41 pages, 59% were less than 8 seconds on the website, 36% were new visitors, and 11% were from México, visited less than 41 pages, navigated at most in the fourth level of the website, and were less than one second on each page.
- From the visitors who logged in, 7% modified a domain, consulted new product info, visited more than 64 pages and were on the webpage more than 31 seconds.
- From the visitors who did not modify a product, 70% did not perform any other conversion, 33% were new visitors that did not consult product availability and were on the website less than 4 seconds, 33% were from México, visited less than 42 pages and navigated at least in the second level of the website.
- From the visitors who modified a product, 16% were recurrent visitors that consulted product A info, logged in, were in the website more than 37 seconds, and did not visit on Sunday.
- 15% were not from México and visited less than 3 pages, 15% were recurrent visitors that consulted product A info, logged in, requested help, visited more than 68 pages, were more than 70 seconds on the website, did not start a payment and did not consult services information.

Bibliography

[1] G2, https://www.g2.com/.

[2] Google analytics - knowledgebase, https://developers.google.com/analytics.

[3] The hypertext transfer protocol (http), https://www.w3.org/protocols/rfc2616/rfc2616.txt.

[4] Log files, http://httpd.apache.org/docs/2.2/logs.htmlcombined.

[5] Matomo - open analytics platform, https://developer.matomo.org.

[6] Omniture website, https://marketing.adobe.com/resources/help.

[7] Public implementation of sequitur in python, https://github.com/markomanninen/pysequitur.

[8] We need to talk about conversion, https://hoteltechreport.com.

[9] Web technology surveys, https://w3techs.com.

[10] Heap website, https://heap.io/blog/product/google-analytics-limits, Jan, 2021.

[11] Leadfeeder website, https://www.leadfeeder.com, Jan, 2021.

[12] Paveai website, https://www.paveai.com/referrer-spam-remover/, Jan, 2021.

[13] Search engine journal website, https://www.searchenginejournal.com/google-analytics-cant-tell/187131/close, Jan, 2021.

[14] Vmo website, https://vwo.com, Jan, 2021.

[15] Woopra website, https://www.woopra.com, Jan, 2021.

[16] ABDELGHANI GUERBAS, OMAR ADDAM, O. Z.-M. N. A. E.-M. R. R. A. Effective web log mining and online navigational pattern prediction. *Knowledge-Based Systems 49 (2013) 50–62* (2013).

[17] AHMED, N. K., AND ROSSI, R. A. Interactive visual graph analytics on the web. *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (2015).

[18] ÁLVARO ROCHA, JOSÉ LUÍS REIS, M. K. P. Z. B. Marketing and smart technologies: Proceedings of icmarktech 2019. *Smart Innovation, Systems and Technologies. Springer Nature. ISSN: 2190-3026* (2019).

[19] ARINDAM BANERJEE, J. G. Clickstream clustering using weighted longest common subsequences. *Dept. of Electrical and Computer Engineering University of Texas* (2020).

[20] ATTA-UR RAHMAN, SUJATA DASH, A. K. L. N. C.-S. B. Y. N. A neuro-fuzzy approach for user behaviour classification and prediction. *Journal of Cloud Computing: Advances, Systems and Applications. 8, 17 (2019). https://doi.org/10.1186/s13677-019-0144-9* (2019).

[21] BÁRBARA CERVANTES, FERNANDO GÓMEZ, O. L.-G. M. A. M.-P. R. M. J. R. Pattern-based and visual analytics for visitor analysis on websites. *Applied Sciences — Open Access Journal* (2019).

[22] BHUPENDRA KUMAR MALVIYA, J. A. A study on web usage mining theory and applications. *2015 Fifth International Conference on Communication Systems and Network Technologies* (2015).

[23] CADEZ IGOR, HECKERMAN DAVID, M. C.-S. P., AND STEVEN, W. Visualization of navigation patterns on a web site using model-based clustering. *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM* (2000).

[24] CHARLESWORTH, A. *Digital Marketing: a Practical Approach*. Taylor and Francis Group. ISBN: 978-0-203-49371-1, 2014. Second edition.

[25] CHRISTOPHE BERTERO, MATTHIEU ROY, C. S.-G. T. Experience report: Log mining using natural language processing and application to anomaly detection. *IEEE 28th International Symposium on Software Reliability Engineering (ISSRE), Toulouse, 2017, pp. 351-360, doi: 10.1109/ISSRE.2017.43* (2017).

[26] CRAIG G. NEVILL-MANNING, I. H. W. Compression and explanation using hierarchical grammars. *The Computer Journal* (1997).

[27] DIA SEKAYI, A. K. Qualitative delphi method: A four round process with a worked example. *The Qualitative Report; Fort Lauderdale Tomo 22, N.º 10, (Oct 2017): 2755-2763.* (2017).

[28] DILEEP KUMAR PADIDEM, C. N. Process mining approach to discover shopping behaviour process model in ecommerce web sites using click stream data. *International Journal of Civil Engineering and Technology (IJCIET)* (2017).

[29] DONG, G. Exploiting the power of group differences: Using patterns to solve data analysis problems. *Synthesis Lectures on Data Mining and Knowledge Discovery 11(1), 1–146 (2019)* (2019).

[30] DOTS, G. 2019 bad bot report. *Machine Vision and Applications* (2019).

[31] DUSAN STEVANOVIC, AIJUN AN, N. V. Feature evaluation for web crawler detection with data mining techniques. *Expert Systems with Applications 39 (2012) 8707–8717* (2012).

[32] FABRICE GUILLET, H. J. H. *Quality Measures in Data Mining*. Springer, ISSN electronic edition: 1860-9503, 2007.

[33] G. NEELIMA, S. R. Predicting user behavior through sessions using the web log mining. *2016 International Conference on Advances in Human Machine Interaction (HMI), Doddaballapur, pp. 1-5, doi: 10.1109/HMI.2016.7449167* (2016).

[34] G. NELSON, J. KIEFFER, P. C. An interesting hierarchical lossless data compression algorithm. *Proceedings 1995 IEEE Information Theory Soc. Workshop, Rydzyna, Poland* (1995).

[35] GALLE, M. Investigating the effectiveness of bpe: The power of shorter sequences. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China* (2019), 1375–1381.

[36] GARY ARMSTRONG, PHILIP T. KOTLER, V. T. L. A. B. *Marketing: An Introduction.* Pearson, 6th edition, ISBN-13: 9780134470528, 2017.

[37] GITA SUKTHANKAR, CHRISTOPHER GEIB, H. H. B. D. P. R. G. *Plan, Activity, and Intent Recognition. Theory and Practice. Chapter 5:Stream Sequence Mining for Human Activity Discovery*. Morgan Kaufmann, 1st. edition, ISBN 9780124017108, 2014.

[38] GÓMEZ, F. Visualization and machine learning techniques to support web traffic analysis. *Thesis of the Master Program in Computer Science at Tecnológico de Monterrey.* (2018).

[39] GRAŻYNA SUCHACKA, J. I. Identifying legitimate web users and bots with different traffic profiles — an information bottleneck approach. *Knowledge-Based Systems 197 (2020) 105875* (2020).

[40] HAROLD A. LINSTONE, M. T. *The Delphi Method, Techniques and Applications*. 2002 Murray Turoff and Harold A. Linstone, 2002.

[41] HUY M. HUYNH, LOAN T. T. NGUYEN, B. V. A. N. V. S. T. Efficient methods for mining weighted clickstream patterns. *Expert Systems with Applications Volume 142, 2020, 112993* (2020).

[42] HUY M. HUYNH, LOAN T. T. NGUYEN, B. V. Z. K. O. T.-P. H. Mining clickstream patterns using idlists. *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Pages 2007–2012, https://doi.org/10.1109/SMC.2019.8914086* (2019).

[43] ISONI, A. *Machine Learning for the Web*. Packt Publishing, 2016.

[44] J. SOONU ARAVINDAN, K. V. An overview of pre-processing techniques in web usage mining. *International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 1 June 2017* (2017).

[45] JAIDEEP SRIVASTAVA, ROBERT COOLEY, M. D. P.-N. T. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations. 1. 12-23* (2000).

[46] JEANNETTE PASCHENA, MATTHEW WILSONB, J. J. Collaborative intelligence: How human and artificial intelligence create value along the b2b sales funnel. *Business Horizons, Volume 63, Issue 3, May–June 202* (2020), 403–414.

[47] JERI JOHN PRABHU DEVAGEORGE (IN), MANIKANDAN VEMBU (IN), S. A. G. U. Methods and systems for grouping and prioritization of website visitors for live support. *U. S. Patent US2020/074519A1, Zoho Corporation Private Limited* (2020).

[48] JIAWEI HAN, MICHELINE KAMBER, J. P. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Series in Data Management Systems. Elsevier Inc. ISBN: 978-0-12-381479-1, 2012, Third Edition.

[49] JOHN C. KIEFFER, E.-H. Y. Lossless data compression via guided approximate bisections. *Conference on Information Sciences and Systems, Princeton University* (2000).

[50] JOSÉ BENITO CAMIÑA, MIGUEL ÁNGEL MEDINA-PÉREZ, R. M.-B. O. L.-G.-L. A. P. V., AND GURROLA, L. C. G. Bagging-randomminer: a one-class classifier for file access-based masquerade detection. *Machine Vision and Applications* (2018).

[51] K. ABIRAMI, P. M. Similarity measurement of web navigation pattern using k-harmonic mean algorithm. *Elysium Journal, Volume-4, Issue-5, October 2017, ISSN: 2347-4734* (2017).

[52] K. ABIRAMI, P. M. Fuzzy clustering with artificial bee colony algorithm using web usage mining. *International Journal of Pure and Applied Mathematics Volume 118 No. 18 2018, 3619-3626, ISSN: 1314-3395* (2018).

[53] K. VELKUMAR, P. T. A survey on web mining techniques. *2nd International Conference on New Scientific Creations in Engineering and Technology (ICNSCET-20) International Journal of Recent Trends in Engineering Research (IJRTER). Special Issue, March 2020. ISSN: 2455-1457* (2020).

[54] LACHLAN A.MAXWELL (US), D. J. M. U. Systems and methods for network traffic analysis. *U. S. Patent US2020/0382542A1, Oath Inc.* (2020).

[55] LATENDRESSE, M. Masquerade detection via customized grammars. *Detection of Intrusions and Malware, and Vulnerability Assessment, Second International Conference (DIMVA) Vienna, Austria, DOI: 10.1007/11506881_9* (2005).

[56] LEONARDO CANETE-SIFUENTES, RAUL MONROY, M. A. M.-P.-O. L.-G.-F. V. V. Classification based on multivariate contrast patterns. *IEEE Access* (2019).

[57] LOYOLA-GONZÁLEZ, O. Supervised classifiers based on emerging patterns for class imbalance problems. *Thesis for the degree of PhD. In Computer Science at INAOE* (2017).

[58] M. SANTHANAKUMAR, C. C. C. Web usage based analysis of web pages using rapidminer. *WSEAS Transactions on Computers* (2015).

[59] MICHAEL, B. *Practical Web Analytics for User Experience*. Elsevier Inc., ISBN: 978-0-12-404619-1, 2013.

[60] MICHELANGELO CECI, P. F. L. Closed sequential pattern mining for sitemap generation. *World Wide Web (2021) 24:175–203 https://doi.org/10.1007/s11280-020-00839-2* (2019).

[61] MICKALA, S. T., AND YOO, Y. User activity recognition in smart homes using pattern clustering applied to temporal ann algorithm. *Department of Electrical and Computer Engineering, Pusan National University* (2015).

[62] MILTON GARCÍA-BORROTO, OCTAVIO LOYOLA-GONZÁLEZ, J. F. M.-T. J. A. C.-O. Evaluation of quality measures for contrast patterns by using unseen objects. *Expert Systems With Applications 83 (2017) 104–113* (2017).

[63] MOSES CHARIKAR, ERIC LEHMAN, D. L. R. P. M. P. A. S. A. S. The smallest grammar problem. *IEEE Transactions on Information Theory Vol. 51*, No. 7 (2005).

[64] MUGHAL, M. J. H. Data mining: Web data mining techniques, tools and algorithms: An overview. *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018* (2018).

[65] MUMTAZ, R. Awareness and perspectives social media as new strategic marketing approach in minor industries; notion grounded on aida model. *Journal of Content, Community Communication, Vol. 10 Year 5, December- 2019, ISSN: 2456-9011* (2019).

[66] N. JESPER LARSSON, A. M. Offline dictionary-based compression. *Proceedings DCC'99 Data Compression Conference (Cat. No. PR00096), doi: 10.1109/DCC.1999.755679* (1999), 296–305.

[67] NEERAJ KANDPAL, H. P. SINGH, M. S. S. Application of web usage mining for administration and improvement of online counseling website. *International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 7 (2019) pp. 1431-1437* (2019).

[68] NIUSVEL ACOSTA-MENDOZA, JESÚS ARIEL CARRASCO-OCHOA, J. F. M.-T. A. G.-A. J. E. M.-P. Mining clique frequent approximate subgraphs from multi- graph collections. *Applied Intelligence (2020) 50:878 –892 https://doi.org/10.1007/s10489-019-01564-8* (2020).

[69] O. EL AISSAOUI, Y. EL MADANI EL ALAMI, L. O.-Y. E. A. Integrating web usage mining for an automatic learner profile detection: A learning styles-based approach. *International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, 2018, pp. 1-6, doi: 10.1109/ISACV.2018.8354021* (2018).

[70] OCTAVIO LOYOLA-GONZÁLEZ, JOSÉ FCO. MARTÍNEZ-TRINIDAD, J. A. C.-O. M. G.-B. Effect of class imbalance on quality measures for contrast patterns: An experimental study. *Information Sciences 374 (2016) 179–192* (2016).

[71] OCTAVIO LOYOLA-GONZÁLEZ, JOSÉ FCO. MARTÍNEZ-TRINIDAD, J. A. C.-O. M. G.-B. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing 175 (Part B) (2016) 935–947* (2016).

[72] OCTAVIO LOYOLA-GONZÁLEZ, MILTON GARCÍA-BORROTO, J. F. M.-T.-J. A. C.-O. An empirical comparison among quality measures for pattern based classifiers. *Intelligent Data Analysis 18 (2014) S5–S17 DOI: 10.3233/IDA-140705* (2014).

[73] OCTAVIO LOYOLA-GONZÁLEZ, M.A. MEDINA-PÉREZ, J. M. J.-R. M. M. G. Pbc4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowledge-Based Systems 115 (2017) 100-109* (2016).

[74] OCTAVIO LOYOLA-GONZÁLEZ, RAÚL MONROY, J. R. A. L.-C. J. I. M.-S. Contrast pattern-based classification for bot detection on twitter. *IEEE Access, April 2019 DOI: 10.1109/ACCESS.2019.2904220* (2019).

[75] OCTAVIO LOYOLA-GONZÁLEZ, MIGUEL ANGEL MEDINA-PÉREZ, K.-K. R. C. A review of supervised classification based on contrast patterns: Applications, trends, and challenges. *J Grid Computing (2020). https://doi.org/10.1007/s10723-020-09526-y* (2020).

[76] OCTAVIO LOYOLA-GONZÁLEZ, RAUL MONROY, M. A. M.-P. B. C. J. E. G.-T. An approach based on contrast patterns for bot detection on web log files. *Wireless Pers Commun (2017) 97:2229–2247* (2017).

[77] OLGA PERDIKAKI, SARAVANAN KESAVAN, J. M. S. Effect of traffic on sales and conversion rates of retail stores. *Manufacturing and Service Operations Management. Vol. 14, No. 1, Winter 2012, pp. 145–162, ISSN 1526-5498* (2012).

[78] P. G. OM PRAKASH, A. J. Analyzing and predicting user navigation pattern from weblogs using modified classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science Vol. 11, No. 1, July 2018, pp. 333 340 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v11.i1. pp. 333-340* (2018).

[79] PABARSKAITE, Z., AND RAUDYS, A. A process of knowledge discovery from web log data: Systematization and critical review. *J Intell Inf Syst (2007) 28:79–104* (2007).

[80] PATRICK M. J. DUBOIS, ZHAO HAN, F. J., AND LEUNG, C. K. An interactive circular visual analytic tool for visualization of web data. *IEEE/WIC/ACM International Conference on Web Intelligence* (2016), 709–712.

[81] PHILIP BILLE, INGE LI GØRTZ, N. P. Space-efficient re-pair compression. *arXiv: 1611.01479* (2016).

[82] PHILIP KOTLER, A. G. *Principles of Marketing*. Pearson Education, 12th edition, ISBN-13: 9780132390026, 2007.

[83] PHILIP KOTLER, HERMAWAN KARTAJAYA, I. S. *Marketing 4.0. Moving from traditional to digital*. John Wiley Sons, Inc. ISBN: 978-1-119-34106-2, 2017.

[84] QIN CHEN, XIANGBIN YAN, T. Z. Converting visitors of physicians' personal websites to customers in online health communities: Longitudinal study. *Journal of Medical Internet Research 2020, vol. 22, iss. 8, http://www.jmir.org/2020/8/e20623/* (2012).

[85] RISHABH MEHROTRA, AHMED EL KHOLY, I. Z. M. S. A. H. Identifying user sessions in interactions with intelligent digital assistants. *Proceedings of the 26th International Conference on World Wide Web Companion. Pages 821–822https://doi.org/10.1145/3041021.3054254* (2017).

[86] ROBERT COOLEY, B. M., AND SRIVASTAVA, J. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems, Springer-Verlag 1999, Volume 1, Issue 1, pp 5–32* (1999).

[87] ROBERT W. PALMATIER, V. KUMAR, C. M. H. *Customer engagement marketing*. Springer Nature. DOI 10.1007/987-3-319-61985-9_1, 2017. Second edition.

[88] RON BERMAN, A. I. The value of descriptive analytics: Evidence from online retailers. *Harvard Business School. Working paper 21-067* (2020).

[89] RYAN DEISS, R. H. *Digital marketing for dummies*. John Wiley Sons, Inc. ISBN: 978-1-119-66049, 2017. Second edition.

[90] RYOSUKE NAKAMURA, SHUNSUKE INENAGA, H. B. T. F. M. T. A. S. Linear-time text compression by longest-first substitution. *Algorithms 2009, ISSN 1999-4893* (2009).

[91] SAPIAN ALINA, V. M. The marketing funnel as an effective way of the business strategy. *Scientific Journal: The art of scientific mind, No. 4, UDC 339.138:004.738.5* (2019), 403–414.

[92] SERIN J., R. L. Clustering based association rule mining to discover user behavioural pattern in web log mining. *International Journal of Pure and Applied Mathematics. Volume 119. No. 17 2018, 1937-1947. ISSN: 1314-3395* (2018).

[93] SHAHIZAN HASSAN, SITI ZALEHA AHMAD NADZIM, N. S. Strategic use of social media for small business based on the aida model. *ScienceDirect, Global Conference on Business Social Science-2014, GCBSS-2014, 15th 16th December, Kuala Lumpur* (2012).

[94] SHRADDHA TIWARI, R. K. G., AND KASHYAP, R. To enhance web response time using agglomerative clustering technique for web navigation recommendation. *Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing 711, doi 978-981-10-8055-5_59* (2019).

[95] STEFANO ROVETTAA, GRAŻYNA SUCHACKAB, F. M. Bot recognition in a web store: An approach based on unsupervised learning. *Journal of Network and Computer Applications 157 (2020) 102577* (2020).

[96] STIELER, M. Creating marketing magic and innovative future marketing trends. *Proceedings of the 2016 Academy of Marketing Science (AMS) Annual Conference. Springer Nature. ISBN: 978-3-319-45596-9* (2017).

[97] SUCHACKA, G. Analysis of aggregated bot and human traffic on e-commerce site. *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1123–1130, ACSIS Vol. 2, DOI: 10.15439/2014F346* (2014).

[98] SUNJYOT SINGH ANAND, ANKIT KUMAR MAMODIA, A. A. K. S. P., AND BHINGARKAR, P. S. A study of classification algorithms for categorizing website users using machine learning. *International Journal of Pure and Applied Mathematics. Volume 118 No. 16 2018, 333-348. ISSN: 1314-3395* (2018).

[99] SUNNY DHAMNANI (IN), VISHWA VINAY (IN), L. K. I. R. S. I. Classification of website sessions using one-class labeling techniques. *U. S. Patent US10,785,318B2, Adobe Inc* (2020).

[100] SVITLANA BONDARENKO, OLENA LABURTSEVA, O. S. V. L. O. H. T. K. Modern lead generation in internet marketing for the development of enterprise potential. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12, October 2019* (2019).

[101] TAN, P.-N., AND KUMAR, V. Discovery of web robot sessions based on their navigational patterns. *Article in Data Mining and Knowledge Discovery, January 2002* (2002).

[102] TAN KAI HUN1, R. Y. The impact of proper marketing communication channels on consumer's behavior and segmentation consumers. *Asian Journal of Business and Management (ISSN: 2321 - 2802) Volume 02 – Issue 02* (2014).

[103] THOMAS SCOTT LEVI (CA), JORDAN TYLER DAWE (CA), Y. S. R.-S. C. M. B. M. C. Method and system for applying machine learning approach to routing webpage traffic based on visitor attributes. *U. S. Patent US2019/024131A1, Unbounce Marketing Solutions Incorporated* (2019).

[104] VENUGOPAL K.R., S. N. S. Web page recommendations based web navigation prediction. in: Web recommendations systems. *Springer, Singapore. https://doi.org/10.1007/978-981-15-2513-1_7* (2020).

[105] VIKAS KUMAR, G. A. O. Web analytics for knowledge creation: A systematic review of tools, techniques, and practices. *International Journal of Cyber Behavior, Psychology and Learning* (2020).

[106] WEICHBROTH, P. Frequent sequence mining in web log data. *5th International Conference on Man–Machine Interactions. In Advances in Intelligent Systems and Computing book series (AISC, volume 659). https://doi.org/10.1007/978-3-319-67792-7_45* (2017).

[107] X. ZHANG, G. D. Overview and analysis of contrast pattern based classification. *Contrast Data Mining: Concepts, Algorithms, and Applications* (2012).

[108] YUANYUAN WANG, HAILIN LIU, Q. L. Application research of web log mining in the e-commerce. *2020 Chinese Control And Decision Conference (CCDC), IEEE, DOI: 10.1109/CCDC49329.2020.9164022* (2020).