IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2020, accepted August 31, 2020, date of publication September 18, 2020, date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024649

Log-Based Session Profiling and Online Behavioral Prediction in E–Commerce Websites

JAVIER FABRA, PEDRO ÁLVAREZ, AND JOAQUÍN EZPELETA
Department of Computer Science and Systems Engineering, Aragón Institute of Engineering Research, Universidad de Zaragoza, 50009 Zaragoza, Spain
Corresponding author: Javier Fabra (jfabra@unizar.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness under Project TIN2017-84796-C2-2-R, and in part by the Aragonese Government under Project DisCo-T21-20R.

ABSTRACT Improvements to customer experience give companies a competitive advantage, as understanding customers' behaviors allows e-commerce companies to enhance their marketing strategies by means of recommendation techniques and the customization of products and services. This is not a simple task, and it becomes more difficult when working with anonymous sessions since no historical information of the user can be applied. In this article, analysis and clustering of the clickstreams of past anonymous sessions are used to synthesize a prediction model based on a neural network. The model allows for prediction of a user's profile after a few clicks of an online anonymous session. This information can be used by the e-commerce's decision system to generate online recommendations and better adapt the offered services to the customer's profile.

INDEX TERMS Behavior prediction, user profiling, log analysis, clustering, neural networks, model checking.

I. INTRODUCTION

E-commerce is a very effective way to bring customers to your business and offer them a 24-7 service. Furthermore, the possibility of keeping the customer connected to the business in a non-face-to-face manner has become a necessity, as reflected by the recent crisis caused by the coronavirus (COVID-19). Analysis of the behaviors and interests of customers is crucial to improve the systems that support e-commerce, with the aim of providing customized services and products to increase the conversion ratio related to purchases [1], [2] and enhance loyalty in certain strategic sectors [3].

Enhancement of a customer's experience gives companies a competitive advantage, as generic marketing makes brands forgettable. The target for a commerce company is loyal and engaged customers who come back and buy again. Prediction of customer behaviors and tastes, as well as individual analysis, is a very complex task that requires a time-consuming integration process. Under a typical configuration, information on the activity of visitors and customers of an e-commerce website is stored in the server logs. To classify the users of the website, clustering techniques are frequently applied to create a segmentation on the available data [4], [5]. Frequently, a series of metrics that revolve around the concept of user sessions have previously been generated. The clusters obtained are then used to define profiles according to the user's browsing history [5]. This history also allows the clusters to be characterized so that it is possible to understand customers' previous actions (transactional data) or their demographic profiles [6]. Nevertheless, it is exponentially more valuable to provide insights concerning what customers will do in the future.

The renewed interest in artificial intelligence techniques has led to a proliferation of methods for predicting the future behavior of e-commerce customers [7], [8]. Most of these methods address the prediction problem from the perspective of aggregated data, providing high-level predictions as a result. Obviously, it is very interesting to know the probability of a customer's purchase during the next visit to the website or whether she/he will be interested in buying a specific product in the future. Nevertheless, the challenge is to make progress towards predictive analytics that offer fine-grained results and increase the dynamism of the business [9], [10]. This new generation of prediction techniques must help adopters to discover potential customers, prevent churn, configure the website's layouts to maximize sales, or offer customized recommendations.

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

Achieving these goals requires that the prediction models be integrated into the company's decision-making systems and that their predictions be validated to adapt those models to the changing conditions of the business and the evolution of customers' habits.

In this article, existing research proposals in the field of customer behavior prediction are reviewed. This review shows the necessity of addressing the challenges towards building fine-grained predictive models and applying them to real scenarios. The research presented in this article advances in this direction by focusing on the behavioral analysis of unregistered customers of an e-commerce, and it shows that it is possible to accurately predict the customer profile of a user session while browsing the website. With respect to existing techniques, the proposed solution addresses the following contributions:

- the prediction model works with incomplete unregistered user sessions;
- the different customers' profiles are explicitly considered as part of the predictions, thus providing a deeper understanding of users' browsing and purchasing behaviors;
- those profiles are interpreted and validated from a business perspective to match predictions with desirable customer behaviors;
- finally, the integration process of a prototype of the solution into a website based on the Magento e-commerce technology is detailed.

The proposal requires building prediction models that are used along with clickstream techniques to analyze the customers' behaviors at runtime. To do that, methodologies for server log processing, clustering algorithms and artificial intelligence techniques are combined in a three-phase process. First, the log files are processed to discover the customer profiles. These profiles are then validated and interpreted from a business perspective, associating each one with a set of behavioral patterns that characterize the users belonging to that profile. After that, a prediction model is created and trained to evaluate users' pattern-based behavior and determine the customer's profile. Alternately, once the models are available, predictions are conducted using the user's clickstream. This allows the system to perform, after a small number of events, precise predictions concerning the segment the session is probably going to fall into.

From this point, the results from the predictions can be used to adapt the customer's session so as to reinforce the prediction or attempt to move the session towards a more interesting segment, according to the e-commerce website's interests. Unlike other existing solutions, predictions are based on the current behaviors of users and not limited by the purchasing probability.

The remainder of this article is organized as follows. Section II focuses on related work. The process for predicting the customer's profile and a real scenario are introduced in Section III. The preprocessing of the log files is presented in Section IV. After that, the clustering process is carried out in Section V. Section VI details the customer profiling and the validation process. The behavior prediction methods used and the obtained results are presented in Section VII. The integration of the prediction system into an e-commerce platform is then detailed in Section VIII. Finally, Section IX outlines some conclusions of this article and addresses future research lines.

II. RELATED WORK

Before making predictions about customers' future behavior, it is necessary to discover the different profiles of users that visit the e-commerce website. The process of profiling consists of two stages: the characterization of customers' past behaviors and the grouping of customers who behave similarly. In this section, the most relevant research approaches related to these two stages will be detailed and analyzed.

Regarding the first stage, most research techniques create customers' behavioral descriptions from the website's log files or the database of customer transactions. The contents of these descriptions can vary depending on the intended use of the analysis results. Customer personal data [11], their RFM (Recency, Frequency and Monetary) values [12]–[16], their browsing behaviors [17], [18] or purchasing habits [19]–[21], or the products they have shown interest in [22]–[25] are typically used for the creation of such descriptions. The concept of session plays a relevant role in this characterization due to the fact that a description is calculated for each customer session. For this reason, the existing approaches are essentially interested in the analysis of registered users. The sessions of these users are clearly identified and directly recorded in the website's log files. As an exception, [22] is the only work dealing with unregistered users. In this case, a process of reconstructing sessions based on clickstream analysis is required [4].

Once customers' descriptions have been created, they are usually grouped using either clustering methods [13], [15], [16], [19], [23], [26] or classification methods [17], [18], [25]. As a result, the application of these techniques generates a set of clusters that must be subsequently interpreted to understand the particular behaviors of each class of customers. An expert-guided analysis of the computed clusters is proposed by some of the approaches [12], [20], [23], [26]. Such a task is rather complicated and time-consuming, and therefore alternatives that automatically extract knowledge from clusters' descriptions should be studied. [11], [19] use association rules for automating these interpretations. Nevertheless, they require advance knowledge of the interesting attributes to define suitable rules. However, the clusters obtained must also be validated. Ideally, the clusters' validation should consist of matching users' future behaviors according to clusters. Some works have proposed frameworks to provide an incremental clustering to dynamically maintain the customer profiles [27], [28]. Despite the efforts to interpret and validate clusters, these are still open challenges.

Customer segmentation techniques are needed to build models that help to make predictions regarding customers' future behaviors. To give a compact view of the existing research in the field of prediction, a set of criteria that help us to classify these works has been established. The result of this classification is presented in Table 1. Let us now detail the classification criteria used.

Prediction goals. Some models address the challenge of distinguishing between buying and non-buying sessions (B/NB, two possible prediction outcomes) [4], [29]–[35], [7], [8]. Alternatively, other works concentrate on calculating the probability that a customer buys either a specific product (B-Prod) [20], [36]–[38] or a class of products (B-CProd) [39], [40], makes a purchase in the next visit to the online store (Next) [41]–[43], or repurchases in a future session (ReP) [44], [45]. Time constraints have also been considered as a part of some prediction models to estimate the purchasing probability of a user for the next day (Next-D) [46], for the next year (Next-Y) [47], or over time (NoT) [40]. Likewise, customers' profiles have been used to distinguish between VIP and non-VIP customers (V/NV) [48]. Notice that the predictions in the cited research require identified users to predict future behavior once the corresponding past behavior has been analyzed.

Data source. Customers' past behavior is usually extracted from log files generated by Web servers (Log-based proposals, Log) or transaction data recorded in the seller's ERP/CRM systems (database approaches, CTD). These data sources are processed to discover and select the features/attributes that will be used to create prediction models. As an exception, [34], [45], [47] propose the use of questionnaires for gathering information regarding customers' preferences and behaviors in the hiring of (banking) services.

Types of customers. Most works make their predictions based on registered customers' past behaviors. Only four works estimate the purchasing probability for unregistered customers [4], [29], [31], [32]. These solutions apply clickstream analysis to reconstruct users' sessions and discover user's behaviors during their navigation through the e-commerce website.

Selection of a predictor. The selection of features is a critical issue for the creation of an accurate prediction model. It consists of extracting/computing a set of relevant attributes from the data source. As shown in Table 1, the most common attributes are customers' personal (P) or demographic (D) data, product interest scores (PI), customers' navigation (NB) or purchasing behaviors (PB), or historical purchasing data (HP) (the RFM value or payments, for instance). Nevertheless, some proposals select alternative interesting attributes, such as the use of shopping carts (SC) [32], [43], seller's reputations and facilities (SRF) [45], customers' opinions (CO) [47], changes in user behavior (ChB) [46] or interactions of users with Web pages and their elements (Int) [7]. From a methodological point of view, two approaches should be emphasized. Firstly, [44] applies feature engineering to automate the selection and ranking of a large number of features to improve prediction tasks. Secondly, a method based on association rules is proposed in [4], [31] as an alternative to traditional techniques.

Prediction techniques. Three types of techniques have been widely applied in the prediction of customers' purchasing behavior: classification methods, regression analysis, and algorithmic techniques. Among the classification methods, the most common are Neuronal Networks (NN), Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF) and Naives-Bayes models (NBM). The approaches based on these methods make their predictions using a single classifier (S-Class, [33], [34], [43], [46]) or by combining multiple classifiers to improve the accuracy of the results (M-Class, [7], [8], [29]–[31], [40], [42], [44], [45]). In the last case, a combination algorithm integrating the predictions of the different classifiers is needed. Genetic algorithms [30] (GA), the Artificial Bee Colony (ABC) algorithm [45], Bootstrap Aggregation (BA) [48] and strategies based on majority voting [20], [31] (MV) have also been used as combination methods. As an exception, [44] assigns weights to models manually (MN).

Alternately, regression analysis has been used to determine the purchasing probability using logistic regression [32], [35], [38], [41], [47]. This statistical model requires an a-priori analysis of the predictors to be used and the correlations between them to make accurate predictions. In [20], a hybrid solution is presented. Logistic regression and classification methods are combined to improve the purchasing predictions of a concrete e-commerce. Although the results notably increase the prediction coverage, prediction accuracy is not clearly improved with respect to other approaches based on the use of a unique technique.

Finally, different algorithms have been proposed to study specific purchasing behaviors [4], [36], [37]. These solutions define probability models that are evaluated in conjunction with association rules to extract the knowledge of interest for e-commerce managers. [36] attempts to discover the most profitable products and customers. It searches for potential customers interested in purchasing a star product in the near future and analyses those buyers' personal profiles. [37] determines the best time (the *peak hour*) for a customer to purchase a product. This time-based information is used to deliver personalized marketing messages to increase sales. [4] uses rules to estimate the purchasing probability of a user session depending on the pages that were visited in the past and the time spent on them.

The nature of approaches. Some of the research works are Application-oriented Approaches (AoA) in the sense that they apply existing prediction methods to solve some concrete problem, usually in the domain of e-commerce or e-banking services. Adopting a different point of view, some works (let us call them Methodology-oriented Approaches, abbreviated as MoA) concentrate on defining new methods/algorithms for predicting future customers' purchasing behaviors. Generally, these types of works also validate

TABLE 1. Comparative analysis of the methods for predicting the purchasing probability.

Reference	Classification Criteria								
	Goals	Data source	Customers	Feature selection	Prediction technique	Nature	Integration	Validation	Extra knowledge
This	CustPf	Log	Unreg	Nav/PB	S-Class(NN)	AoA	✔	✗	Expert
[4]	B/NB	Log	Unreg	Nav/PB	Apriori Alg.	AoA	✗	✗	AsR
[29]	B/NB	Log	Unreg	Nav/PB	S-Class(SVM)	AoA	✗	✗	✗
[48]	V/NV	CTD	Reg	P/RFM/HP	M-Class(BA)	AoA	✗	✗	AsR
[30]	B/NB	CTD	Reg	D/HP	M-Class(GA)	M/AoA	✗	✗	✗
[34]	B/NB	Qt	Reg	P/HP	S-Class(NN)	AoA	✗	✗	✗
[36]	B-Prod	Log	Reg	P/Nav/PB	Pred-Alg	M/AoA	✗	✗	AsR
[37]	B-Prod	CTD	Reg	RFM	Pred-Alg	AoA	✗	✗	✗
[47]	Next-Y	Qt	Reg	CO/HP	Regression	AoA	✗	✗	BSM
[31]	B/NB	Log	Unreg	Nav	M-Class(MV)	AoA	✗	✗	AsR
[41]	Next	CTD/Log	Reg	D/Nav/PB	Regression	AoA	✗	✗	✗
[33]	B/NB	Log	Reg	Nav/HP	S-Class(NBM)	AoA	✗	✗	✗
[46]	Next-D	Log	Reg	P/Nav/HP/PI	S-Class(DT)	AoA	✗	✗	✗
[38]	B-Prod	Log	Reg	PI	Regression	AoA	✗	✗	✗
[42]	Next	CTD	Reg	HP	M-Class	M/AoA	✗	✗	✗
[43]	Next	Log	Reg	D/HP/SC	SVM-Alg	M/AoA	✗	✗	✗
[45]	ReP	Qt	Reg	PI/HP/SRF	M-Class(ABC)	AoA	✗	✗	✗
[20]	B-Prod	CTD	Reg	HP	Hybrid	AoA	✗	✗	✗
[32]	B/NB	CTD	Unreg	NB/PB/SC	Regression	AoA	✗	✗	✗
[35]	B/NB	Log	Reg	NB/PB	Regression	AoA	✗	✗	✗
[44]	ReP	Log	Reg	P/D/PI/NB/PB/SC	M-Class(MN)	AoA	✗	✗	✗
[8]	B/NB	Log	Reg	NB/PB	M-Class(NN)	AoA	✗	✗	✗
[7]	B/NB	Log	Reg	NB/PB/Int	M-Class(NN)	AoA	✔	✗	✗
[40]	B-CProd/NoT	Log	Reg	NB/PB	M-Class(NN)	AoA	✗	✗	✗
[39]	B-CProd	CTD	Reg	D/HP	PGM	M/AoA	✗	✗	✗

their solutions by applying them to real application cases (MoA/AoA).

Integration into e-commerce websites. Prediction methods help explain customers' behaviors. This understanding can be used to improve the design and contents of websites, perform various recommendation techniques, increase the effectiveness of marketing campaigns, or customize the service for the user, for instance. In spite of these possibilities, most solutions have not been integrated into a real e-commerce system, except for [7], which developed a prototype of a system that can be installed on users' mobile devices. Therefore, the integration of the predictions in the lifecycle of e-commerce websites is an open challenge that should be addressed.

Validation of results by experts. Prediction models and algorithms are usually trained and tested using the data recorded in server logs or transaction databases. Moreover, different metrics (recall, precision, etc.) have been defined to evaluate the quality of the predictions. Nevertheless, other supplementary methods should be applied to evaluate the real usefulness of predictions to create new business value and opportunities. These methods could consist of a qualitative validation of results based on expert opinions, for instance. Alternately, because the prediction models are not usually integrated in real systems, the predictions are not validated with customers' future behaviors. As an exception, [47] assesses the validation of predictions (the probability that a customer buys during the next year) by comparing them with the purchases made during the year following the publication of the paper.

Discovery of extra knowledge. Many of the proposals aim to classify customers' behaviors. Nevertheless, the reasons that lead customers to exhibit that behavior are not usually studied. Some works apply association rules (AsR) to analyze the sessions with high purchasing probability to discover behavioral patterns and the reasons that lead to the purchase of some products [4], [31], [36], [48]. The knowledge discovered is limited and consists of simple relationships between pairs of navigation/purchasing events. As an alternative, [47] builds a Behavioral Scoring Model (BSC). These models have been widely used to identify frequent user behaviors in the field of financial services, but their applicability to online commerce must still be investigated.

III. A PROCESS FOR PREDICTING CUSTOMERS' PROFILES

Our goal is to create a model that helps to predict the possible future behavior of a customer session while browsing an e-commerce website. This prediction can be used to influence customers' actions (for example, to improve purchase intentions and/or probability) or to provide them with customized contents or products. The proposed approach consists of applying a process in three phases: preprocessing of the server log files, discovery of the customer profiles, and synthesis of the behavioral model. The final result is a prediction model that is integrated into the e-commerce's decision system to personalize the services it offers to customers.

The process followed in this article is similar to that used in [49] in the field of predictive business process monitoring. In that case, a two-phase approach analyzes incomplete traces of business processes to predict at runtime whether their execution outcomes will be as expected. These predictions help to minimize the likelihood of violation of business constraints specified using Linear Temporal Logic (LTL). That technique is applied over medical processes with a well-defined structure. It is a relevant difference with respect

to our approach, in which a user can navigate freely through the website's structure.

A. BUILDING A PREDICTION MODEL

Figure 1 shows the three-phase process followed in this article. Firstly, the raw e-commerce logs are preprocessed to discard uninteresting requests, identify user sessions and prepare the log contents to enable their analysis. The result of the preprocessing phase is a collection of sessions. A session is an ordered sequence of user interactions with the system (events) that take place within a time frame. A session can contain multiple page views, events, social interactions and e-commerce transactions corresponding to actions such as visiting a page, executing a search, adding/deleting a product to/from the cart or completing the payment process, for instance. A session can be interpreted in terms of users' behaviors. The process is designed to be useful for systems with either logged or anonymous access. In the first case, a session is clearly established in terms of a sequence of events corresponding to the logged session. In the case of anonymous access, a *sessionization* process is required to establish which events in a sequence can be considered as belonging to the same session, as will be detailed later.

Afterwards, the established sessions are used to determine the e-commerce's *customer profiles*. This second phase starts by computing a *vector of features* for each session (the *features creation* task). A feature provides a high-level and (usually) quantitative description of the user's behavior during the session: the total session time, the number and type of visited pages, or the number of times that the resources were used (the search engine, the cart, the wishlist, etc.), for instance. In the case of logged e-commerce websites, features can be enhanced with demographic and geographic data, buying patterns of previous user sessions, or the purchase history, for instance. The features are then processed by the *clustering* task to group in the same cluster those sessions that present similar features (that are assumed to be strongly related to the user's behavior). These first two tasks can be executed several times to improve the features' expressivity and to find a more adequate (optimal) number of clusters as well.

Once the session clusters have been computed, they are interpreted from a business perspective. The business analyst is responsible for mapping these clusters to customer profiles. This *profile discovery* is complex and requires knowledge of the website's structure, the customer's interests and purchase habits, and the types of users that typically interact with the e-commerce. Finally, the resulting profiles must be validated. This task consists of checking if the behavior of a cluster's session corresponds to the behavioral description of its profile. To do that, some type of study of the conformance of the sessions in a cluster and the intuitive description established for it must be carried out. For instance, if a cluster is described as corresponding to spurious website users, one can expect those sessions to enter the system from a different point than the home URL (maybe coming from an external search engine) and also be short sessions.

The aim of the *prediction model synthesis* phase is to generate a prediction model so as to be able to establish, after a few session events, the cluster to which a live session is probably going to belong. The inputs for this phase are the set of clusters and some behavioral indexes for each session that typically are associated with initial stages of the session. The prediction model will be used to analyze the event stream of each session and, after the considered initial stage, predict the cluster of the considered session. Different (artificial intelligence or statistical) techniques can be applied to obtain the prediction model. In this work, neural network pattern recognition techniques are applied during the process, but it could be easily adapted to different alternatives. For that, a vector of features is obtained for the first k events of each session. The features and clusters feed an artificial neural network *synthesis method*, which is trained and validated (in the *Model training* and *Quality analysis* tasks, respectively). The resulting artificial neural network is the model used to predict the session behavior.

Finally, the model obtained is integrated into the e-commerce's prediction system. E-commerce data logs are processed during the customer's navigation and transformed into events of interest. These events represent the customer's actions during the browsing (visiting a product category or the product itself, using the search engine, adding/deleting a product to/from the cart, or completing the purchase, among others). The prediction system interprets the event stream and determines the most probable customer profile, which will be used by the e-commerce system to make some decisions or recommendations adapted to the session behavior.

B. THE UP&SCRAP USE CASE

The process presented to obtain a prediction model will be applied over a real e-commerce website. Specifically, it will be applied over the website of Up&Scrap[1], a scrapbooking company with more than 25, 000 clients all around the world. In this subsection, the structure and contents of this website are introduced.

The structure of the website of Up&Scrap is organized around two different types of sections (main and secondary). Each section is then split into several subsections to refine the product classification. Figure 2 depicts the structure of the website. Similar taxonomies have been proposed by different authors but including only main sections [23], [50]. From the homepage (level 0), different sections can be accessed (level 1). Two different types of sections can be distinguished. *Main sections* organize products according to their functionality and utility. The website provides a menu to access this main categorization of products. There are eight different sections (papers, decorations, stamps, tools, project life-smash, albums, home decor-DIY, and gifts), which are divided into subcategories. Alternately, there are *secondary*

[1] http://www.upandscrap.com

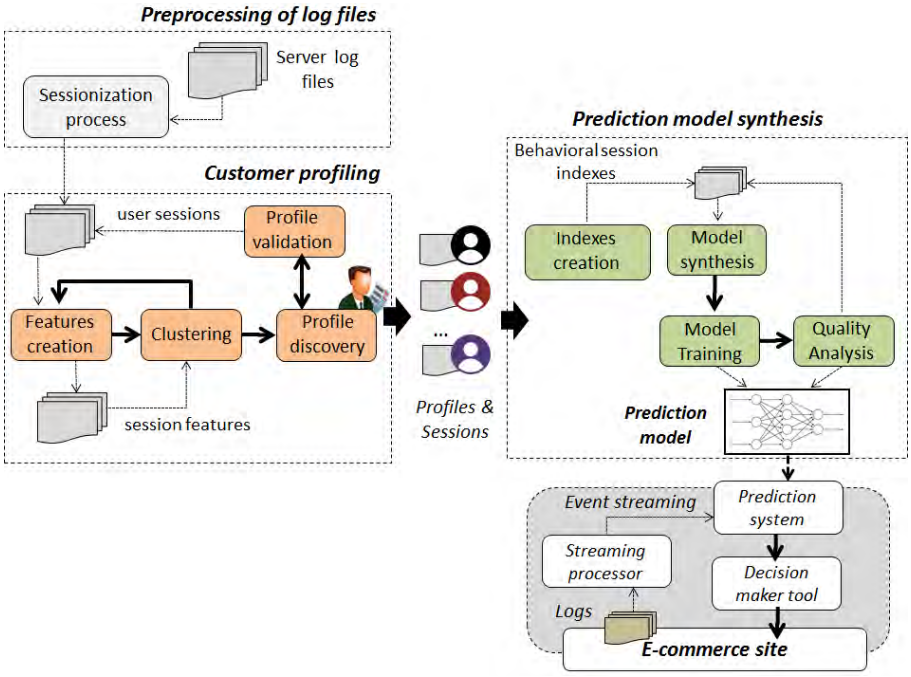

FIGURE 1. Sketch of the process for identifying and predicting the customers' profiles.

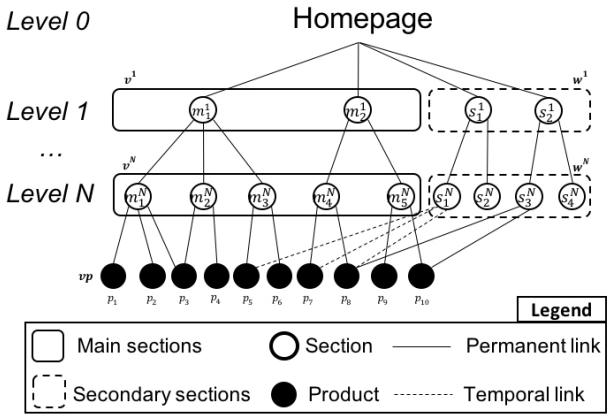

FIGURE 2. Abstract view of the structure of the Up&Scrap website.

sections, which classify products according to other complementary criteria, such as designers, themes/collections, brands, offers, or new products, for instance. The secondary menu provides access to six different secondary sections. The website also includes a search engine that allows users to directly look for products without using the proposed sections. As a consequence, the website structure allows users to reach products following many different navigation paths, as well as offering direct access via the search engine.

In the following, let us detail the process of identifying and predicting the customers' profiles for Up&Scrap through the different phases depicted in Figure 1.

IV. PREPROCESSING OF LOG FILES

The website logs follow the Common Language Format standard (CLF) [51] and provide raw information such as the IP address from which the session was established, the timestamp of the request, the page URL, or the HTTP status returned to the client, for instance. The data that is being handled in this article correspond to the analysis of the log provided by the system corresponding to two months of use, and 8, 607, 625 events are contained. The log corresponds to non-logged users sessions, and there is no user info. Figure 3 shows a piece of the raw web log of the considered scenario.

First, a cleaning and filtering process to remove undesired and uninteresting records for the behavioral analysis of users was applied. Specifically, events corresponding to the following criteria were removed: automatic requests, such as the ones performed by robots, spiders and crawlers; requests with erroneous status codes that are not relevant for navigational patterns; requests of irrelevant HTTP methods (only GET and POST requests have been considered because they are unique requests directly from users); and finally, requests asking for multimedia contents automatically generated by the browser. After these steps, the log was reduced to 5, 875, 479 records, 68.26% of the original size.

After that, a log preparation stage was carried out. The aim of this process is to prepare the log file for the clustering process. Two types of actions are performed to this end. In the *categorization* sub-phase, each record is analyzed to identify high-level events and extract meaningful information. Additionally, log contents are reduced in the *simplification* sub-phase to increase the effectiveness of the post-processing.

During the log categorization, different events can be identified by analyzing the CLF log contents. For such purpose, each event is automatically classified by considering

1.2.4.5 - - [14/Apr/2019:02:21:41 +0100] "GET /estampar HTTP/1.1" 200 28883 "https://www.upandscrap.com" "Mozilla/5.0 (iPhone; CPU iPhone OS 10_3 like Mac OS X) AppleWebKit/602.1.50 (KHTML, like Gecko) CriOS/56.0.2924.75 Mobile/14E5239e Safari/602.1" "-" "0.776"
3.3.4.1 - - [14/Apr/2019:02:21:43 +0100] "GET /colecciones/perlitas HTTP/1.1" 200 19231 "https://www.upandscrap.com/colecciones" "Mozilla/5.0 (Linux; Android 5.1.1; Nexus 5 Build/LMY48B; wv) AppleWebKit/537.36 (KHTML, like Gecko) Version/4.0 Chrome/43.0.2357.65 Mobile Safari/537.36" "-" "0.754"
1.2.4.5 - - [14/Apr/2019:02:21:49 +0100] "GET /estampar/tintas HTTP/1.1" 200 28945 "https://www.upandscrap.com/papeles" "Mozilla/5.0 (iPhone; CPU iPhone OS 10_3 like Mac OS X) AppleWebKit/602.1.50 (KHTML, like Gecko) CriOS/56.0.2924.75 ”Mobile/14E5239e Safari/602.1" "-" "0.792"
1.2.4.5 - - [14/Apr/2019:02:22:02 +0100] "POST /checkout/cart/add/uenc/b7J29H2M8dJs9L27Z82Y9201UuBG89H2,/product/44293/form_key/c521f290b23a27e8/ HTTP/1.1" 200 128 "https://www.upandscrap.com/ estampar/tintas" "Mozilla/5.0 (iPhone; CPU iPhone OS 10_3 like Mac OS X) AppleWebKit/602.1.50 (KHTML, like Gecko) CriOS/56.0.2924.75 Mobile/14E5239e Safari/602.1" "-" "0.701"
3.3.4.1 - - [14/Apr/2019:02:21:05 +0100] "GET /colecciones HTTP/1.1" 200 28534 "https://www.upandscrap.com/colecciones/distress-crayons" "Mozilla/5.0 (Linux; Android 5.1.1; Nexus 5 Build/LMY48B; wv) AppleWebKit/537.36 (KHTML, like Gecko) Version/4.0 Chrome/43.0.2357.65 Mobile Safari/537.36" "-" "0.812"

FIGURE 3. An extract from the raw log of the web server, where IP addresses have been anonymized.

whether it is a GET or POST request and analyzing its URL in terms of the presence and/or absence of specific keywords and resources (that is, the words between slash characters). These requests are classified based on their deepness and whether they correspond to a main or a secondary section. In the case of the Up&Scrap website, its structure is organized in two levels ($N = 2$). The different events have then been separated into 63 different types, such as *Visit main section L1, Visit secondary section L2*[2]*, Visit product, Login, Logout, Add product to the wishlist, Add product to the cart*, etc. These events refer to different actions and can affect different sections of the website. However, not all event types are interesting for the analysis because some of them provide superfluous information. They can, for instance, refer to user account management or legal warnings.

Therefore, in the simplification stage, some of these event types are discarded, and only the event types that are interesting for the type of analysis that is going to be conducted are considered, with the aim of reducing the amount of information included in the log by filtering the records that do not contain relevant information. To do that, the following filters are applied. First, sessions with fewer than three requests are discarded because they do not contain valuable information and mainly correspond to users that do not have an interest in the website contents.

Second, some events are discarded because they do not provide valuable information for the analysis. Because the goal of analyzing the logs is to extract information regarding users' behaviors and preferences when buying products, there are many events that can be considered superfluous, such as events related to user account management or rating products. In this case, a set of 12 types of events that considered relevant for the analysis has been identified, and the remaining ones have been filtered. In the following, they are detailed according to the different sets identified: *Visit_main_section_L1*, *Visit_secondary_section_L1* (accessing the results of the search engine is considered to be visiting a secondary section), *Visit_main_section_L2*, *Visit_secondary_section_L2*, *Visit_homepage* (an event type representing that the homepage has been visited), *Visit_product* (an event type representing that the URL of a product has been visited), *Add_wishlist_products_to_the_cart*, *Add_product_to_the_cart*, *Add_product_to_the_wishlist*, *Buy_products_in_the_cart*, *Delete_product_from_the_cart*, and finally, *Update_product_ from_the_cart*.

The last filter removes duplicated events, which reduced the log size to $1,331,697$ records. Figure 4 shows the processed log of the extract depicted in Figure 3. A more detailed description of the entire process can be found in [52].

After that, a sessionization process to group those events that could be considered as belonging to the same session was conducted; because this study deals with non-logged sessions, additional criteria had to be applied to define the start and end events of each session. For that, a session as the ordered sequence of events from the same IP for which no more than 30 minutes passed between any two consecutive events was defined. This is a common characterization that has been used in log file analysis to discover knowledge by several authors [53]–[55]. As a result, $138,085$ anonymous sessions were identified.

V. CLUSTERING PROCESS

The next step consists of the clustering of sessions with some common characteristics. To do that, for each session, a set of global properties was extracted. The set of properties that can be interesting is strongly related to the problem domain. In the domain of this work, it has to be dependent on the website structure because the structure will constrain the types of sequences of events a user can execute. For the structure in Figure 2, the properties described in Table 2 have been considered.

For each session, the set of corresponding values is used to generate the *vector of features*. This vector is useful for providing a high-level view of the session (abstracting

[2]Note that L1 and L2 are related to the two levels of the Up&Scrap website

Id	IP	Timestamp	Event name	Relative URL	Operation	Code	L1 section	L2 section
1	1.2.4.5	14/Apr/2019:02:21:41 +0100 +0100	Visit main section L1	/estampar	GET	200	estampar	
1	1.2.4.5	14/Apr/2019:02:21:49 +0100	Visit main section L2	/estampar/tintas	GET	200	estampar	tintas
1	1.2.4.5	14/Apr/2019:02:22:02 +0100	Add product to the cart	/checkout/cart/add/...	POST	200		
2	3.3.4.1	14/Apr/2019:02:21:43 +0100 +0100	Visit secondary section L2	colecciones/perlitas	GET	200	colecciones	perlitas
2	3.3.4.1	14/Apr/2019:02:21:05 +0100	Visit secondary section L1	/colecciones	GET	200	colecciones	

FIGURE 4. Log generated after the preprocessing phase from the web server log's extract depicted in Figure 3.

TABLE 2. Properties considered as characterizing user sessions.

	Attribute	Meaning
1	*LONG*	# of events that compose the trace representing the session
2	*DUR*	session duration, in seconds
3	*TMED*	mean time duration of events, in seconds
4	*HOME*	# of visits to the homepage
5	*ML1*	# of visits to elements belonging to the first level of main category
6	*SL1*	# of visits to elements belonging to the second level of main category
7	*ML2*	# of visits to elements belonging to the first level of secondary category
8	*SL2*	# of visits to elements belonging to the second level of secondary category
9	*CART*	# of operations that add an item to the cart
10	*WHISH*	# of items added to the whishlist
11	*COUT*	# of checkout operations
12	*SEARCH*	# of search operations
13	*OFFER*	# of operations related to marketing campaigns and offers
14	*NOV*	# of operations related new products
15	*PROD*	# of views of the detail page of products

unimportant details) and facilitating their interpretation by the business analyst. Of the 15 properties identified in Table 2, 9 are left, which are those most relevant for prediction issues. Of these 9 properties, some can be grouped/added for the clustering process. The other properties that have not been used in this phase can be useful and interesting later to perform certain validation processes. Specifically, the following features have been considered: *MAIN*, which groups visits to the main category ($MAIN = ML_1 + ML_2$); *SECONDARY*, which groups visits to the secondary category ($SECONDARY = SL_1 + SL_2$); *MARKETING*, which groups marketing-related events ($MARKETING = OFFER + NOV$); *INTEREST*, which groups events that indicate interest ($INTEREST = WISH + PROD + CART$); and finally, the *SEARCH* feature, which corresponds to the property of the same name. As a result, a feature for each session is obtained. As the next step, the sessions are clustered.

As the clustering technique, k-Means has been applied to the vector of features. The objective of this algorithm is to partition a set of n elements into k groups of "near" elements: each element belongs to the group whose average is closer. The algorithm requires definition of the number of clusters, k. To find the optimal value of k, Knime [56] and R [57] were used.

In Knime, a workflow that performs an iterative process and calculates the entropy generated by the selection of different values of k has been developed. Entropy is a measure of the variation of the attributes in the data set for each cluster; the closer the value is to 0, the greater the similarity of the data is. However, the further away from 0, the greater differences between data were (and thus worse results are obtained during the clustering process). A lower entropy determines the optimal number of clusters in which the data should be grouped. In the case of the R software, the NbClust package [58] was used. This package provides 30 indexes to determine the optimal number of clusters and proposes the best grouping scheme based on the different results obtained. This includes very well-known methods such as Silhouette or Pamk (which uses the PAM or Clara algorithms, along with the Silhouette method).

The clustering process was conducted using the sessions whose lengths were greater than one event, feeding the process with the described features. One-event sessions can be considered as noise traces. From the original dataset, which contained 138, 085 sessions, there are 101,917 traces whose lengths are longer than one event ($LONG > 1$). Both Knime and R provided us with the same optimal number of clusters, $k = 4$.

Table 3 provides information regarding the results of the clustering and the mean values for the features of each cluster with respect to the considered sessions. Each element in the table corresponds to the mean value in the considered set. For instance, the normalized global mean value of the MAIN value is 0.3538, while that constrained to Cluster 1 is 0.0348. Cluster 1 contains 20, 273 sessions (19.9% of the total number of sessions), cluster 2 contains 38, 670 sessions (37.9%), cluster 3 contains 15,573 sessions (15.3%), and finally, cluster 4 contains the remaining 27, 401 sessions (26.9%).

The analysis of the features of each cluster with respect to the initial set of data shows that there is a set of feature values that stands out for each cluster (the values have

TABLE 3. Normalized clustering results for $k = 4$.

Feature	Full data	Cluster 1	Cluster 2	Cluster 3	Cluster 4
MAIN	0.3538	0.0348	**0.6640**	0.1492	0.2685
SECONDARY	0.2535	**0.6594**	0.0507	**0.4810**	0.1100
MARKETING	0.0772	**0.3165**	0.0121	0.0168	0.0265
INTEREST	0.2499	0.1464	0.1635	0.1685	**0.4948**
SEARCH	0.0758	0.0133	0.0254	**0.3559**	0.0338
# Traces	101,917	20,273	38,670	15,573	27,401

been highlighted in bold in Table 3), which can be used to establish the users' profiles. Cluster 1 shows normalized mean values of 260% for the SECONDARY and 409% for the MARKETING attributes, respectively. Cluster 2 shows a ratio of visits to the main page (MAIN) of 189% with respect to the global average. Cluster 4 stands out in the SECONDARY and SEARCH values, with averages of 189% and 469% with respect to the global average, respectively. The fact that the search events are part of the secondary ones is clearly reflected in the correlated values of such attributes in this cluster. Finally, cluster 4 shows that INTEREST stands out with values that represent 197% with respect to the global average. There are other coincidences regarding the outstanding features among the clusters, which will help in the characterization process.

VI. CUSTOMER PROFILING AND VALIDATION

The analysis of the data obtained from the clustering process allows us to perform an initial profiling phase [59]. Customer profiling is the subdivision of a market into discrete customer groups that share similar characteristics [60], [61]. This process allows for identification of common characteristics among different users and potential customers, as well as the proposal of retargeting strategies. Customer profiling requires, as key steps, the division of the market into meaningful and measurable segments (clusters) according to customers' needs, past behaviors or demographic profiles (if available), as well as determination of the profit potential of each cluster by analyzing those aspects and characteristics that stand out in each one.

A. CLUSTER INTERPRETATION

Let us now perform a cluster interpretation focusing on the salient features of the clusters obtained in the previous section, as well as the results shown in Table 3. In addition, the information that the clusters offer us allows for the addition of certain characteristics based on the properties calculated previously. The values that stand out above the others for each feature appear in bold in Table 3.

As shown, cluster 1 stands out in the features of MARKETING and secondary sections (SECONDARY). These values indicate that this cluster groups the customers who usually access (or repeat their visit to) the website via a campaign or marketing source (which correspond to secondary items). Those customers also focus on secondary items (SECONDARY feature) and do not tend to move out of the visited category or explore among different categories, instead remaining in a narrow navigation area. They have a low browsing dispersion. Analyzing session duration, it was found that these customers' sessions are short.

The second cluster stands out in the MAIN property, which indicates that the users falling into this cluster represent first-time customers or customers that spend time browsing the website. They probably are users that land on the website for exploratory or purchasing purposes. These users have long-term sessions with high dispersion (low focus on the same level/category of items), and they show a moderate ratio of purchases.

A detailed view of cluster 3 emphasizes that, as in cluster 1, the secondary-section property stands out. However, this cluster also highlights the search engine property (SEARCH), which allows intuiting that the population that falls into cluster 3 corresponds to those customers that browse using the search engine (search-based navigation). There are three main options that probably explain this behavior: possible ignorance of the website map, the aim of looking for very specific items, or a non-specific purchasing/browsing focus. Additionally, depending on the session duration, two different subclasses of customers with this profile can be distinguished: one with sessions that have low time between events, which represent customers that usually will not finally purchase; and another one with very specific customers who visit the product page to finally purchase it. These last are represented by sessions that are characterized with a longer time between events.

Finally, customers grouped in cluster 4 stand out for the INTEREST property, which groups the wishlist, visits to detail pages of the product, and actions in the shopping cart. This indicates that these customers may have a clear idea about the website and the products and categories they are interested in. It can be observed also that these customers focus more on main sections than on secondary ones, visit more product pages, and spend some time on them. Usually, they enter through the homepage, and their sessions end with just purchasing products or keeping products in the cart. Sessions belonging to this profile are long and concentrated on interest-related events.

Based on the main characteristics of each cluster, let us provide a named classification. This helps to create a conceptual separation among the groups, similarly to other approaches [62]:

- customers in cluster 1 correspond to *repeat or geek customers*;
- customers in cluster 2 correspond to *explorer customers*;

- customers in cluster 3 correspond to *searcher customers* (or narrow searcher customers for very specific ones); and finally,
- customers in cluster 4 correspond to *potential or prospective buyers*.

B. CLUSTER VALIDATION

The complete data (features along with the initial properties) from the clustering process was used to validate the clusters using model checking techniques [63]. To this end, this study proposes a set of validation queries related to results obtained from the clustering process.

Events have been defined as propositional variables and grouped into formulas, as depicted in Table 4, for better understanding. The meaning of every event can be easily deduced from the event name. Alternately, &, | and ! correspond, respectively, to *and*, *or* and *not* logic connectives.

Some of these queries are detailed below in natural language:

- How do users access the website (Q_1: How many sessions directly access the website through a *MARKETING* event?).
- How do users access main sections (Q_2: How many sessions visit neither the main L1 nor main L2 sections?).
- Which is the relation between how do users access the website and purchasing (Q_3: How many sessions access through a *MARKETING* event and then have a *PURCHASE* event?).
- How do users use the search engine of the website (Q_4: How many sessions never use the search engine? $Q_{5,6}$: How many sessions use the search engine at least three (Q_5) or four times (Q_6)? Q_7: How many sessions feature intensive use of the search engine (at least 4 times) and purchase items?).
- How do customers purchase products (Q_8: How many sessions have a *PURCHASE* event? Q_9: How many sessions operate with the cart and then checkout? Q_{10}: How many sessions have two or more purchase events?)
- What is the relation between user navigation and purchases (Q_{11}: How many sessions iterate at least five times between main and secondary sections and do not purchase in the end? Q_{12}: How many sessions visit five or more product pages?)
- How do users purchase and iterate with the cart (Q_{13}: How many sessions add products to the cart but do not purchase in the end? Q_{14}: How many sessions add at least two products to the cart? $Q_{15,16}$: How many sessions have two *PURCHASE* events, but the cart remains untouched (Q_{15})/is modified (Q_{16}) between the checkouts? Q_{17}: How many sessions contain three or more *PURCHASE* events?)

Table 5 shows the results obtained for each cluster (c_i), reflecting the percentage of sessions that the validation query (Q_j) fulfills with respect to the total number of sessions of the cluster, as well as the queries described in the LTL version proposed in [63]. In the table, operators *G,F*, and *X* have the usual LTL interpretation: *Always*, *Eventually* and *Next*, respectively, being *H,O*, and *Y* in past counterparts. Alternately, *x,y,z*, appearing in queries Q_{15} and Q_{16}, correspond to *freeze* operators, allowing us to talk about specific positions in the session and providing the capacity to relate attributes of different session events.

The answers to the questions are different depending on the clusters. The characteristics that stand out for a cluster with respect to the others have been highlighted in green in Table 5, while those less prominent but equally important features have been highlighted in yellow.

As it is shown, queries Q_1 and Q_2 especially highlight sessions in cluster 1, corresponding to the repeat or geek customers according to the initial interpretation. This indicates that the marketing campaigns mainly target this type of client (Q_1, 24%) and that they mainly access the secondary sessions of the website (Q_2, 85%). Query Q_2 also shows that the searcher customers (cluster 3) have a high percentage (44%) centered on the secondary sections. This makes sense because the use of the search engine allows for refinement of the navigation on the website, giving direct access to brands and products (which are located in the secondary categories). In addition, query Q_2 indicates that the explorer customers (cluster 2) always visit the main categories at both the L_1 and L_2 levels, which means that customers within this profile *explore* the website through these more general categories.

The third query (Q_3) tells us the impact that marketing campaigns have on purchases. As can be seen, the percentages are very low and only highlight two clusters, cluster 4 with 3% of sessions that end up buying from marketing campaigns and, on the contrary, cluster 1, where there are no sessions in which a marketing campaign produces a purchase. This has a direct relationship with the initial interpretation of the clusters because cluster 4 corresponds to a buyer client profile (hence the highest percentage), while cluster 1 represents users who access the website, especially to browse secondary sections, but without a buyer profile.

Queries Q_4 through Q_7 allow us to study the use of the search engine of the website and its relationship with purchases. Query Q_4 tells us that the sessions of clusters 1, 2 and 4 use the search engine not very often (between 93% and 71% of the sessions never use it), while in the sessions of cluster 3, the search engine is an event that does appear very frequently (92% of the sessions). This behavior verifies that the sessions in cluster 3 correspond to a searcher customer profile. Queries Q_5 and Q_6 allow us to go into detail regarding those sessions that use the search engine, noting that a high percentage (Q_6, 28%) use it at least four times during a session. Finally, the query Q_7 allows us to see the relationship between the use of the search engine and the purchase events. As shown in the table, the low percentages obtained (2% and 3% for clusters 3 and 4, respectively) indicate that the use of the search engine does not have an evident impact on the purchasing processes.

TABLE 4. Definition of formulas used in validation queries.

Formula	Definition
MAIN_Q	ev_Visit_main_section_L1 | ev_Visit_main_section_L2
SECONDARY_Q	ev_Visit_secondary_section_L1 | ev_Visit_secondary_section_L2
MARKETING_Q	ev_Campaigns | ev_New_products
INTEREST_Q	ev_Add_product_to_the_wishlist | ev_Visit_product | ev_Add_product_to_the_cart
SEARCH_Q	ev_Catalog_search
BUY_Q	ev_Buy_products_in_the_cart

TABLE 5. Results for the profile's validation process.

Query	LTL	C_1: Repeat/Geek customers	C_2: Explorer customers	C_3: Searcher customer	C_4: Potential buyers (prospective)
Q_1	MARKETING_Q	24%	2%	3%	4%
Q_2	!(F MAIN_Q))	85%	0%	44%	24%
Q_3	(F MARKETING_Q) & (F BUY_Q)	0%	1%	1%	3%
Q_4	G !SEARCH_Q	93%	81%	8%	71%
Q_5	F (SEARCH_Q & X F (SEARCH_Q & X F SEARCH_Q))	1%	2%	6%	3%
Q_6	F (SEARCH_Q & X F (SEARCH_Q & X F (SEARCH_Q & X F SEARCH_Q)))	0%	2%	28%	8%
Q_7	(F BUY_Q) & F (SEARCH_Q & X F (SEARCH_Q & X F (SEARCH_Q & X F SEARCH_Q)))	0%	0%	2%	3%
Q_8	F BUY_Q	0%	3%	4%	13%
Q_9	F (ev_Add_product_to_the_cart & X F BUY_Q)	0%	3%	3%	11%
Q_{10}	F (BUY_Q & X F BUY_Q)	0%	2%	2%	8%
Q_{11}	G(!BUY_Q) & F(MAIN_Q & X F (MAIN_Q & X F(MAIN_Q & X F (MAIN_Q & X F MAIN_Q))))	25%	46%	40%	39%
Q_{12}	F(ev_Visit_product & X F (ev_Visit_product & X F(ev_Visit_product & X F (ev_Visit_product & X F ev_Visit_product))))	6%	15%	10%	38%
Q_{13}	(F ev_Add_product_to_the_cart) & !(F BUY_Q)	6%	11%	12%	28%
Q_{14}	F (ev_Add_product_to_the_cart & X F ev_Add_product_to_the_cart) & ! (F BUY_Q)	3%	6%	7%	21%
Q_{15}	F x.(BUY_Q & X F (BUY_Q & H y.((y[#]-x[#]<0) | ((!ev_Add_product_to_the_cart) & (!ev_Delete_product_from_the_cart)))))	0%	2%	2%	7%
Q_{16}	F x.(BUY_Q & X F y.(BUY_Q & O z.((ev_Add_product_to_the_cart | ev_Delete_product_from_the_cart) & (z[#]-x[#]>0) & (y[#]-z[#]>0))))	0%	0%	0%	2%
Q_{17}	F (BUY_Q & X F (BUY_Q & X F BUY_Q))	0%	2%	2%	8%

Queries Q_8 through Q_{17} allow us to study the purchase process and its relationship with other events. Query Q_8 tells us where the sessions that make purchases are at some point. As can be observed, sessions in cluster 1 (repeat or geek customers) never make a purchase on the website. Customers from the other profiles do make purchases, but where they concentrate is in cluster 4 (13% vs. 3% and 4% of clusters 2 and 3, respectively), which effectively corresponds to potential/prospective buyers. The same behavior is repeated in queries Q_9 and Q_{10}, which confirms the results obtained.

Queries Q_{11} and Q_{12} are very interesting to elucidate the *intention* of the clients of the clusters. Query Q_{11} allows us to observe that cluster-2 sessions stand out for navigating between main and secondary sections but do not end up buying (46%). In addition, in this cluster, there is a significant number of sessions that perform at least five views of specific products (Q_{12}, 15%). This corresponds to an explorer profile, which validates our initial hypothesis that the cluster-2 sessions correspond to the explorer customers. Alternately, query Q_{12} gives us more information about the buyers (cluster 4): a high percentage (38%) of the sessions have at least five visits to the products, which is natural in a purchase profile.

Queries Q_{13} through Q_{17} are oriented toward the interaction with the cart. As shown, the greatest interaction with the cart occurs in the sessions of cluster 4 (Q_{13}, 28%; Q_{14}, 21%), where a purchase is not made in the end. This explains why the profile of cluster 4 includes prospective buyers: they are buyers who have not finished deciding but who probably end up buying the products in the next session. Queries Q_{15} and Q_{16} allow us to study what happens between two purchase events: as can be seen, there is a significant percentage (Q_{15}, 7%) of sessions in cluster 4 in which the customer does not modify the cart between one checkout and another.

This indicates that the client started the checkout, decided to go back to review a product, but finally ended up not modifying the content of the cart. It is very interesting to be able to observe this behavior from anonymous sessions where there is no information associated with the detail of the checkout process.

Finally, query Q_{17} confirms that cluster 4 corresponds to buyers (or potential buyers) because 8% of their sessions make three purchase events at some time. The percentages in the other clusters are much lower or nonexistent (2% in clusters 2 and 3, and no sessions in cluster 1).

This process of validating the clusters through the use of queries with temporal logic allows us to validate the initial hypothesis the clusters obtained, as well as providing additional information (use of website components such as the search engine, impact of marketing campaigns, details of the purchase process, etc.) that can be very valuable for the business expert.

From the validation, it should be proven that the four clusters correspond to the interpretation given in the previous subsection.

VII. BEHAVIOR PREDICTION

Previous sections have established the interest in grouping users' behaviors by means of clustering techniques. At this point, a few interesting questions concerning the relationship between clusters and the users' behaviors appear. Is it possible to predict the cluster to which a session will belong to by analyzing a few initial events? If so, how many events are required to get a good prediction? How accurate is that prediction?

Answering the previous questions is useful when one wants to modify the user's behavior to reach some desired objectives. Let us consider, for instance, the case in which very few users of a given cluster buy products. Predicting the case after a few events is essential to apply recommendation policies with the aim of redirecting the user session towards a different cluster, one more related to the searched objective. On the contrary, if the up-to-now user behavior predicts that the session is going to belong to a cluster strongly related with buyers, the interest will be in ensuring that she or he does not abandon the behavior associated with that cluster.

Table 6 shows some global data regarding the clusters and also the relations between clusters and buying sessions. Each row corresponds to a cluster. The columns correspond, respectively, to the number of sessions, the percentage of sessions with respect to the total number of sessions in the log, the number of buying sessions, the percentage of buying sessions with respect to the total number of sessions in the log, and finally, the percentage of buying sessions with respect to the total number of buying sessions. Notice that most of the buying sessions are concentrated in clusters 4 (64.81%) and 2 (23.63%). Let us consider that, after a few initial events of a session (5, for instance), the system is able to detect that the session is very likely going to belong to cluster 1 or 3. This means that it will be quite probable that the user will buy no product, and then some suggestions could be proposed to drive him towards one of the clusters with a larger buying probability (namely, 2 or 4).

TABLE 6. Some data about sessions and clusters.

	# sessions	% wrt sessions	# buying	% wrt #sessions	% wrt # buying
C1	20,273	18.89	96	0.09	1.70
C2	38,670	37.94	1,330	1.30	23.63
C3	15,573	15.28	554	0.54	9.84
C4	27,401	26.88	3,648	3.58	64.81
Total	101,917		5,628	5.52	

With the aim of correlating initial user behavior and clusters, in this study some machine learning techniques have been applied. The process carried out is as follows:

- First, a vector of features for each session has been computed. The features are based on the same values used for clustering (MAIN, SECONDARY, MARKETING, INTEREST and SEARCH, as described in Section V), but the counting of event occurrences is constrained to the first n events, with n varying from 3 to 8. Different prediction models using the first 3 to 8 events are going to be obtained.
- As the second step, a multilayer feed-forward network has been built. It is composed of 5 hidden layers, with 10 hidden neurons per layer. The learning algorithm applied is RProp [64], constraining its execution to up to 100 learning iterations. The network has been trained with the scaled conjugate gradient back-propagation method using 75% of randomly selected sessions (76,478 session features).
- Finally, the remaining 25% sessions (25,479 session features) have been used to test the quality of the resulting pattern recognition method, whose results are commented on in the following.

The results obtained are summarized in different tables. Table 7 corresponds to the confusion matrices of the prediction models based on 3 (left) and 4 events (right); Table 8 corresponds to the use of 5 and 6 events; and Table 9 corresponds to the cases of 7 and 8 events. As an example, let us describe the case of 5 events (left part of Table 8).

Rows $AC1$ through $AC4$ correspond to the clusters to which input sessions belong (*actual clusters*), while columns $PC1$ through $PC4$ correspond to the predicted clusters according to the trained neural network (*predicted clusters*). Concentrating on a row, the diagonal element corresponds to the correctly predicted sessions, while the rest are false negatives (id est, input features that should be predicted as belonging to the cluster corresponding to the row but that have been predicted as belonging to a different one). For instance, row 2 in Table 8 (left) shows that 8,238 (true positives) cluster-2 sessions where properly predicted as belonging to that cluster, while 269, 254 and 906 cluster-2 sessions where predicted as belonging to clusters 1, 3 and 4, respectively (false negatives). The value in the *Rec.* column

TABLE 7. Confusion matrices for the test phase when computing the feature for the 3 (left) and 4 (right) first events.

	PC1	PC2	PC3	PC4	Rec.	PC1	PC2	PC3	PC4	Rec.
AC1	4558	251	127	132	0.90	4758	32	46	232	0.94
AC2	377	7600	538	1152	0.79	351	8095	201	1020	0.84
AC3	452	391	2750	301	0.71	457	360	2662	415	0.68
AC4	875	1451	688	3836	0.56	658	1343	376	4473	0.65
Prec.	0.73	0.78	0.67	0.71		0.76	0.82	0.81	0.73	

TABLE 8. Confusion matrices for the test phase when computing the feature for the 5 (left) and 6 (right) first events.

	PC1	PC2	PC3	PC4	Rec.	PC1	PC2	PC3	PC4	Rec.
AC1	4838	38	70	122	0.95	4837	45	44	142	0.95
AC2	249	8383	329	706	0.87	138	8744	154	631	0.90
AC3	393	236	3046	219	0.78	371	287	3005	231	0.77
AC4	583	1305	430	4532	0.66	452	1196	343	4859	0.71
Prec.	0.80	0.84	0.79	0.81		0.83	0.85	0.85	0.83	

TABLE 9. Confusion matrices for the test phase when computing the feature for the 7 (left) and 8 (right) first events.

	PC1	PC2	PC3	PC4	Rec.	PC1	PC2	PC3	PC4	Rec.
AC1	4881	17	58	112	0.96	4852	12	48	156	0.96
AC2	112	8687	248	620	0.90	69	8908	122	568	0.92
AC3	312	132	3254	196	0.84	312	166	3168	248	0.81
AC4	418	998	362	5072	0.74	319	1047	269	5215	0.76
Prec.	0.85	0.88	0.83	0.85		0.87	0.88	0.88	0.84	

corresponds to what is called the *recall* value, as a measure of the quality of the prediction for the considered cluster, and is computed with the following formula:

$$\text{recall} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false negatives}}$$

Let us now concentrate on columns. Column values out of the diagonal correspond to false positives: sessions predicted as belonging to the cluster associated with the column while actually belonging to a different cluster. For instance, considering column 2 of the same table, 38, 236 and 1,305 sessions where predicted as belonging to cluster 2 when they actually belonged to clusters 1, 3 and 4, respectively. The value in the *Prec.* row corresponding to what is called *precision*, which is computed with the following formula:

$$\text{precision} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false positives}}$$

Precision and recall provide insight into the prediction quality for each cluster. To measure the global quality, *accuracy* and *Cohen's kappa* statistics are typically considered. Accuracy is computed, for the entirety of the data, as:

$$\text{accuracy} = \frac{\#\text{true positives}}{\#\text{instances}}$$

Accuracy provides an intuitive global view of the quality of the predictions. *Cohen's kappa* is used to measure to what degree two different systems of prediction are in agreement. In this case, it is used to compare the accuracy of the prediction system (observed accuracy) with respect to the accuracy of a random system (expected accuracy) [65], [66]. The kappa value has been computed according to the following formula [66]:

$$\mathcal{K} = \frac{N \cdot \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i\cdot} \cdot x_{\cdot i}}{N^2 - \sum_{i=1}^{k} x_{i\cdot} \cdot x_{\cdot i}}$$

where x_{ii} is the number of cases in the i position of the main diagonal, $N = 25,479$ is the number of sessions, $k = 4$ is the number of clusters, and $x_{\cdot i}, x_{i\cdot}$ are the total number of sessions in the i-th column and row, respectively.

TABLE 10. *Accuracy* and *Cohen's kappa* values for the predictions based on 3 to 8 events.

#events	3	4	5	6	7	8
Accuracy	0.74	0.78	0.82	0.84	0.86	0.87
Cohen's kappa	0.64	0.70	0.75	0.78	0.81	0.82

Table 10 shows, for the different models, the values of the *accuracy* and *Cohen's kappa* statistics. Depending on the authors and the problem domain, there are different scales dividing the kappa value domain, from *non-agreement* to *almost perfect agreement*. What values of the kappa statistic are interesting? There are different interpretations. [65] established negative values as indicating that there is no agreement, 0.01-0.20 as having little agreement, 0.21-0.40 as fair agreement, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement. Alternately, [67] considers that scale to be unacceptable for some domains (in healthcare research, for instance), proposing an alternative scale: 0-0.20 as no agreement, 0.21-0.39 as minimal, 0.40-0.59 as weak, 0.60-0.79 as moderate, 0.80-0.90 as strong, and above 0.9 as almost perfect.

As was expectable, the quality of the prediction improves when more initial events are considered for the prediction.

The online analysis system can make a first initial prediction as soon as a minimal number of initial events have occurred (for example, 4) and then pass the information to the component in charge of applying some recommendation policies. If a new event occurs, the prediction with the corresponding model can be reported, and so on.

VIII. INTEGRATION OF THE PREDICTION SYSTEM INTO AN E-COMMERCE WEBSITE

Finally, let us depict the integration process of a prototype of the solution into the Up&Scrap website and discuss its practical implications from a business perspective.

A. DESCRIPTION OF THE TECHNICAL PROPOSAL

The *Up&Scrap* website was developed using the *Magento Commerce Platform* [68]. Magento provides sophisticated functionality to create customized and secured e-commerce websites, analyze the business to accelerate the sales, and manage an enterprise's product catalog, inventory or marketing channels, among others. Moreover, its open-source nature has encouraged the development of a wide variety of extensions and themes that help programmers to improve website capabilities and presentation, respectively. These features have led Magento to become one of the most popular solutions on the market.

The goal is now to improve customers' shopping experience by predicting their future behavior. This requires the integration of the new prediction system based on the previously computed models into the backend of the *Up&Scrap* website. A reusable system design is proposed to favor its future integration into other Magento websites. The extensible architecture and the technological stack of Magento have helped us to address this design issue.

Figure 5 shows the high-level design of the proposed prediction system. On the left part of the figure, an abstraction of the initial Magento-based e-commerce system for *Up&Scrap* is represented. It consists of a customized instance of the layered framework of Magento 2.0 (colored in orange). This version of Magento allows us to store data logs about customers' navigation through the website, system's internal processes and application server's performance. These logs comply with the PSR-3 standard and were generated and stored as files by using *Monolog*, a standard logging library for PHP [69].

Monolog allows the programming of advanced logging strategies. A new *Monolog handler* is required. This handler is responsible for registering customers' navigation and actions through the website and sending the corresponding log messages to the preprocessing server via a socket connection. This logging handler is integrated into the backend's presentation layer. The role of the preprocessing server is similar to the sessionization component described in Figure1, that is, it discards uninteresting requests and identifies user sessions. Nevertheless, in this case, the messages of interest related to a session will be created progressively as the customer is browsing the website. Then, *Apache Kafka* [70] is used to process these messages, organizing them into (incomplete) sessions and generating an event stream that describes what is happening in each of these in-process sessions.

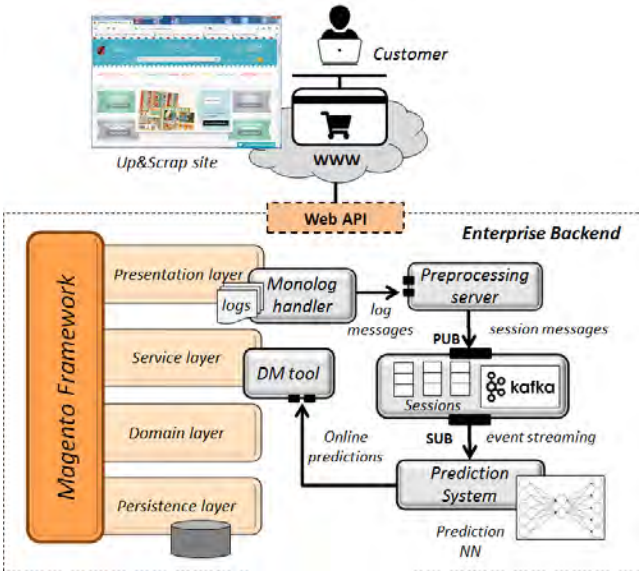

FIGURE 5. Integration of the prediction system into a Magento e-commerce website.

The neural network built for predicting the clusters to which sessions belong is integrated into the *prediction system*, which is subscribed to the Kafka event streaming and uses these events to run the trained neural network. The prediction consists of determining to which cluster a session is more likely to belong based on the incomplete information available thus far. These online predictions are then sent to a *decision-maker tool*, developed as a Magento extensible module that offers its functionality as a service contract. This design decision facilitates its integration into the service layer of the backend's e-commerce system. Internally, predictions can be used to provide Web content that changes based on the behavior, preferences, and interests of the customer or to send personalized offers/discounts by e-mail to customers during their browsing, among things.

B. PRACTICAL IMPLICATIONS OF THE INTEGRATION

The introduction of changes in a business must be a progressive process oriented towards consolidation of the improvements. In the case of Up&Scrap, the results of the customer profiling and the prediction system have been used to understand the current state of its e-business and propose a roadmap of changes that increases the sales. This roadmap has been structured in three phases.

The beginning phase is based on the knowledge extracted from the customer profiles. In this phase, the business experts are mainly interested in corroborating their intuitions about Up&Scrap customers' behaviors and adopting corrective actions that improve the organization of the contents and the

navigation structure of the e-commerce website. Therefore, the changes are directed to improve the user experience during the e-shopping. The analytics tools integrated into the current version of the e-commerce provide high-level insights regarding the customers' navigation habits, but they were not useful for performing fine-grained analysis of those behaviors. The techniques applied to the creation and validation of profiles have been demonstrated to be useful to extract those low-level insights from the server logs. Moreover, these techniques can be reused to validate the results of the changes proposed, analyzing the logs stored after those changes.

In the second phase, the results of the prediction system are used to turn the Up&Scrap e-commerce website into a dynamic application able to offer a more personalized service to the customers. The interest of the company is specially focused on improving the mechanisms of online marketing as a way of influencing the customers' behavior during their navigation. The predictions are interpreted to determine the contents that are introduced in the banners and pop-up messages shown dynamically to each customer. These contents provide feedback to the user about the products, offers and services (for instance, workshops) that could be of interest to her/him. The advantage of this customized marketing strategy is that the e-commerce system does not require significant technical changes. Moreover, these types of improvements are also applied to the marketing by e-mail, sending more personalized product recommendations and offers that induce the customers to buy in future sessions.

Finally, the third phase is the most ambitious and involves a change in the Up&Scrap technological infrastructure. The goal is for the e-commerce system to dynamically adapt its contents and navigation structure to each customer during the shopping. These adaptations would be based on the results of the prediction system and directed to maximize the probability that the customer buys in that session. Unfortunately, the Up&Scrap e-commerce (and any solution based on Magento or other similar technologies) has a static nature that does not allow modification of its contents and structure at runtime and with the desired flexibility. Therefore, this phase would involve the development of a new version of the e-commerce system and a set of base technologies that support the dynamism required.

IX. CONCLUSION

This article has concentrated on the prediction of users' behaviors in e-commerce websites. First, session traces have been grouped according to the similarity of a set of quantitative session parameters. After that, the resulting values for each cluster have been compared with the entire log dataset to define a user profile for each cluster. The profiles have been validated (or refined) by a closer inspection of the sessions in clusters so as to confirm or contradict the (intuitive) cluster profiling description. A later training phase has generated the prediction model that will be used for the online prediction. As a result, the cluster to which the session is going to belong can be predicted after a few initial session events.

Despite the process being applicable to a wide domain, each step requires specific adaptations when considering its application to specific cases: in the preprocessing phase, where some events are discarded for different reasons that cause events to be considered as non-user events, or non-interesting events, for instance; in the clustering phase, where the vector of features as well as an adequate number of clusters must be established; in the profiling phase, where the ratios between global and cluster values are chosen to define the cluster profiles; in the validation phase, where the cluster-profile relations are evaluated and, perhaps, changed; and in the prediction model synthesis phase, where the number of initial events is chosen as an adequate parameter to obtain an accurate online prediction. Every taken decision is arguable. However, some of the steps are easily automatized. Most clustering tools are able to find an optimal number of clusters, and this is also the case when looking for an adequate number of initial events for prediction, for instance.

Further research is required, and different techniques could be applied and adopted for the clustering and prediction model synthesis phases. With respect to the vector of features, the attributes considered are mainly quantitative but do not consider causal relations among events in a sequence. For instance, when counting the number of times two events, a and b, appear in a (partial) session, the possibility of a always appearing after b, or vice versa, is not distinguished, which could hint at very different behaviors, thus corresponding to different profiles. In this sense, the inclusion in the features of such types of relations (or more complex ones) using the answers to temporal logic formulas describing such relations, for instance, could be a way of improving the results.

REFERENCES

[1] R. Kohavi, "Mining E-commerce data: The good, the bad, and the ugly," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2001, pp. 8–13.

[2] N. Verma and J. Singh, "An intelligent approach to big data analytics for sustainable retail environment using apriori-MapReduce framework," *Ind. Manage. Data Syst.*, vol. 117, no. 7, pp. 1503–1520, Aug. 2017.

[3] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019.

[4] G. Suchacka and G. Chodak, "Using association rules to assess purchase probability in online stores," *Inf. Syst. e-Bus. Manage.*, vol. 15, no. 3, pp. 751–780, 2017.

[5] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *Electron. Commerce Res. Appl.*, vol. 14, no. 1, pp. 1–13, Jan. 2015.

[6] R. E. Bucklin and C. Sismeiro, "Click here for Internet insight: Advances in clickstream data analysis in marketing," *J. Interact. Marketing*, vol. 23, no. 1, pp. 35–48, Feb. 2009.

[7] L. Guo, L. Hua, R. Jia, B. Zhao, X. Wang, and B. Cui, "Buying or browsing?: Predicting real-time purchasing intent using attention-based deep network with multiple behavior," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1984–1992.

[8] D. Koehn, S. Lessmann, and M. Schaal, "Predicting online shopping behaviour from clickstream data using deep learning," *Expert Syst. Appl.*, vol. 150, Jul. 2020, Art. no. 113342.

[9] K. Močarníková and M. Greguš, "Conceptualization of predictive analytics by literature review," in *Data-Centric Business and Applications* (Lecture Notes on Data Engineering and Communications Technologies), vol. 30, N. Kryvinska and M. Greguš, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-19069-9_8.

[10] J. Chen and A. Abdul, "A session-based customer preference learning method by using the gated recurrent units with attention function," *IEEE Access*, vol. 7, pp. 17750–17759, 2019.

[11] W. Niyagas, A. Srivihok, and S. Kitisin, "Clustering e-banking customer using data mining and marketing segmentation," *ECTI Trans. Comput. Inf. Technol.*, vol. 2, no. 1, pp. 63–69, Jan. 1970.

[12] H. C. C. Chai, "Online auction customer segmentation using a neural network model," *Int. J. Appl. Sci. Eng.*, vol. 3, no. 2, pp. 101–110, 2005.

[13] M. Namvar, M. R. Gholamian, and S. KhakAbi, "A two phase clustering method for intelligent customer segmentation," in *Proc. Int. Conf. Intell. Syst., Model. Simulation*, Jan. 2010, pp. 215–219.

[14] L. B. Romdhane, N. Fadhel, and B. Ayeb, "An efficient approach for building customer profiles from business data," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1573–1585, Mar. 2010.

[15] M. Walters and J. Bekker, "Customer super-profiling demonstrator to enable efficient targeting in marketing campaigns," *South Afr. J. Ind. Eng.*, vol. 28, no. 3, pp. 113–127, Nov. 2017.

[16] A. Beheshtian-Ardakani, M. Fathian, and M. Gholamian, "A novel model for product bundling and direct marketing in e-commerce based on market segmentation," *Decis. Sci. Lett.*, vol. 7, no. 1, pp. 39–54, 2018.

[17] J. X. Yu, Y. Ou, C. Zhang, and S. Zhang, "Identifying interesting customers through Web log classification," *IEEE Intell. Syst.*, vol. 20, no. 3, pp. 55–59, May 2005.

[18] P.-H. Chou, P.-H. Li, K.-K. Chen, and M.-J. Wu, "Integrating Web mining and neural network for personalized e-commerce automatic service," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 2898–2910, Apr. 2010.

[19] J. Wilson, S. Chaudhury, and B. Lall, "Clustering short temporal behaviour sequences for customer segmentation using LDA," *Expert Syst.*, vol. 35, no. 3, Jun. 2018, Art. no. e12250.

[20] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: A case study," *Marketing Intell. Planning*, vol. 35, no. 4, pp. 544–559, Jun. 2017.

[21] M. R. Flores-Méndez, M. Postigo-Boix, J. L. Melús-Moreno, and B. Stiller, "A model for the mobile market based on customers profile to analyze the churning process," *Wireless Netw.*, vol. 24, no. 2, pp. 409–422, Feb. 2018.

[22] I. S. Y. Kwan, J. Fong, and H. K. Wong, "An e-customer behavior model with online analytical mining for Internet marketing planning," *Decis. Support Syst.*, vol. 41, no. 1, pp. 189–204, Nov. 2005.

[23] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *Electron. Commerce Res. Appl.*, vol. 14, no. 1, pp. 1–13, Jan. 2015.

[24] S. Dhaliwal, N. N. Van, M. Dhaliwal, J. Rokne, R. Alhajj, and T. Ozyer, "Integrating SOM and fuzzy k-means clustering for customer classification in personalized recommendation system for non-text based transactional data," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*, May 2017, pp. 901–908.

[25] S. Palaniappan, A. Mustapha, C. F. Mohd Foozy, and R. Atan, "Customer profiling using classification approach for bank telemarketing," *Int. J. Inform. Vis.*, vol. 1, nos. 4–2, p. 214, Nov. 2017.

[26] K. Kalaidopoulou, S. Triantafyllou, A. Griva, and K. Pramatari, "Identifying customer satisfaction patterns via data mining: The case of greek e-shops," in *Proc. 11th Medit. Conf. Inf. Syst. (MCIS)*, 2017, pp. 1–12.

[27] C. Haruechaiyasak, C. Tipnoe, S. Kongyoung, C. Damrongrat, and N. Angkawattanawit, "A dynamic framework for maintaining customer profiles in E-commerce recommender systems," in *Proc. IEEE Int. Conf. e-Technol., e-Commerce e-Service*, Mar. 2005, pp. 768–771.

[28] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain, "A Web usage mining framework for mining evolving user profiles in dynamic Web sites," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 2, pp. 202–215, Feb. 2008.

[29] G. Suchacka, M. Skolimowska-Kulig, and A. Potempa, "Classification of E-Customer sessions based on support vector machine," in *Proc. Eur. Council Modelling Simulation (ECMS)*, May 2015, 594–600.

[30] E. Kim, W. Kim, and Y. Lee, "Combination of multiple classifiers for the customer's purchase behavior prediction," *Decis. Support Syst.*, vol. 34, no. 2, pp. 167–175, Jan. 2003.

[31] E. Suh, S. Lim, H. Hwang, and S. Kim, "A prediction model for the purchase probability of anonymous customers to support real time Web marketing: A case study," *Expert Syst. Appl.*, vol. 27, no. 2, pp. 245–255, Aug. 2004.

[32] J. Qiu, Z. Lin, and Y. Li, "Predicting customer purchase behavior in the e-commerce context," *Electron. Commerce Res.*, vol. 15, no. 4, pp. 427–452, Dec. 2015.

[33] R. Jia, R. Li, M. Yu, and S. Wang, "E-commerce purchase prediction approach by user behavior data," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2017, pp. 1–5.

[34] L. M. Badea, "Predicting consumer behavior with artificial neural networks," *Procedia Econ. Finance*, vol. 15, pp. 238–246, Jan. 2014.

[35] M. Zeng, H. Cao, M. Chen, and Y. Li, "User behaviour modeling, recommendations, and purchase prediction during shopping festivals," *Electron. Markets*, vol. 29, no. 2, pp. 263–274, Jun. 2019.

[36] H.-J. Chang, L.-P. Hung, and C.-L. Ho, "An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis," *Expert Syst. Appl.*, vol. 32, no. 3, pp. 753–764, Apr. 2007.

[37] N. Vanessa and A. Japutra, "Contextual marketing based on customer buying pattern in grocery E-Commerce: The case of Bigbasket. com (India)," *ASEAN Marketing J.*, vol. 9, no. 1, pp. 56–67, 2018.

[38] N. Nishimura, N. Sukegawa, Y. Takano, and J. Iwanaga, "A latent-class model for estimating product-choice probabilities from clickstream data," *Inf. Sci.*, vol. 429, pp. 406–420, Mar. 2018.

[39] Y.-T. Wen, P.-W. Yeh, T.-H. Tsai, W.-C. Peng, and H.-H. Shuai, "Customer purchase behavior prediction from payment datasets," in *Proc. 11th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2018, pp. 628–636.

[40] C. Huang, X. Wu, X. Zhang, C. Zhang, J. Zhao, D. Yin, and N. V. Chawla, "Online purchase prediction via multi-scale modeling of behavior dynamics," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2613–2622.

[41] D. Van den Poel and W. Buckinx, "Predicting online-purchasing behaviour," *Eur. J. Oper. Res.*, vol. 166, no. 2, pp. 557–575, Oct. 2005.

[42] K. Shapoval and T. Setzer, "Next-purchase prediction using projections of discounted purchasing sequences," *Bus. Inf. Syst. Eng.*, vol. 60, no. 2, pp. 151–166, Apr. 2018.

[43] J. Li, L. Tang, A. Wang, and Z. Xu, "Online-purchasing behavior forecasting with a firefly algorithm-based SVM model considering shopping cart use," *EURASIA J. Math., Sci. Technol. Educ.*, vol. 13, no. 12, Nov. 2017, 7967–7983.

[44] G. Liu, T. T. Nguyen, G. Zhao, W. Zha, J. Yang, J. Cao, M. Wu, P. Zhao, and W. Chen, "Repeat buyer prediction for E-Commerce," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 155–164.

[45] A. Kumar, G. Kabra, E. K. Mussada, M. K. Dash, and P. S. Rana, "Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention," *Neural Comput. Appl.*, vol. 31, no. S2, pp. 877–890, Feb. 2019.

[46] D. Li, G. Zhao, Z. Wang, W. Ma, and Y. Liu, "A method of purchase prediction based on user behavior log," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1031–1039.

[47] K. K. Boyer and G. T. M. Hult, "Customer behavior in an online ordering application: A decision scoring model," *Decis. Sci.*, vol. 36, no. 4, pp. 569–598, Nov. 2005.

[48] B. Shim, K. Choi, and Y. Suh, "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7736–7742, Jul. 2012.

[49] C. D. Francescomarino, M. Dumas, F. M. Maggi, and I. Teinemaa, "Clustering-based predictive process monitoring," *IEEE Trans. Services Comput.*, vol. 12, no. 6, pp. 896–909, Nov. 2019.

[50] Y. S. Kim and B.-J. Yum, "Recommender system based on click stream data using association rule mining," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13320–13327, Sep. 2011.

[51] Common Log Format (CLF). (1995). *The World Wide Web Consortium (W3C)*. [Online]. Available: http://www.w3.org/Daemon/User/Config/Logging.html#common-logfileforma

[52] S. Hernandez, P. Alvarez, J. Fabra, and J. Ezpeleta, "Analysis of Users' behavior in structured e-Commerce websites," *IEEE Access*, vol. 5, pp. 11941–11958, 2017.

[53] *Google Analytics Help Center*. Accessed: Sep. 2020. [Online]. Available: https://support.google.com/analytics

[54] G. Suchacka and G. Chodak, "Practical aspects of log file analysis for E-commerce," in *Proc. Int. Conf. Comput. Netw.*, 2013, pp. 562–572.

[55] M. Adnan, M. Nagi, K. Kianmehr, R. Tahboub, M. Ridley, and J. Rokne, "Promoting where, when and what? An analysis of Web logs by integrating data mining and social network techniques to guide ecommerce business promotions," *Social Netw. Anal. Mining*, vol. 1, no. 3, pp. 173–185, Jul. 2011.

[56] *KNIME*. Accessed: Sep. 2020. [Online]. Available: https://www.knime.com
[57] *The R Project for Statistical Computing*. Accessed: Sep. 2020. [Online]. Available: https://www.r-project.org
[58] *NbClust: Determining the Best Number of Clusters in a Data Set*. Accessed: Sep. 2020. [Online]. Available: https://cran.r-project.org/web/packages/NbClust/index.html
[59] R.-S. Wu and P.-H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," *Electron. Commerce Res. Appl.*, vol. 10, no. 3, pp. 331–341, May 2011.
[60] Bain&Company. (2018). *Management Tools-Customer Segmentation*. [Online]. Available: https://www.bain.com/insights/management-tools-customer-segmentation
[61] J. S. E. Almquist and N. Bloch, "The elements of value," in *Harvard Business Review*. Brighton, MA, USA: Harvard Business Publishing, 2016.
[62] Business2community. (2016). *The Ultimate Guide to eCommerce Customer Segmentation*. [Online]. Available: https://www.business2community.com/ecommerce/ultimate-guide-ecommerce-customer-segmentation-01624275
[63] J. M. Couvreur and J. Ezpeleta, "A linear temporal logic model checking method over finite words with correlated transition attributes," in *Data-Driven Process Discovery and Analysis. SIMPDA* (Lecture Notes in Business Information Processing), vol. 340, P. Ceravolo, M. van Keulen, and K. Stoffel, Eds. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-11638-5_5.
[64] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 1, Mar. 1993, pp. 586–591.
[65] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
[66] A. Bendavid, "Comparison of classification accuracy using Cohen's weighted kappa," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 825–832, Feb. 2008.
[67] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, pp. 276–282, Oct. 2012.
[68] (2020). *Magento E-commerce Platform*. [Online]. Available: https://magento.com/
[69] *Monolog: Sends Your Logs to Files, Sockets, Inboxes, Databases and Various Web Services*. Accessed: Sep. 2020. [Online]. Available: https://packagist.org/packages/monolog/monolog
[70] (2020). *Apache Kafka: A Distributed Streaming Platform*. [Online]. Available: https://kafka.apache.org/

JAVIER FABRA received the Ph.D. degree in computer science from the University of Zaragoza, Spain, in 2010. He has been an Associate Professor with the Department of Computer Science and Systems Engineering, University of Zaragoza, Spain, since 2008. His main research interests include data mining analysis techniques in the context of service-oriented computing and cloud architectures.

PEDRO ÁLVAREZ received the Ph.D. degree in computer science engineering from the University of Zaragoza, Zaragoza, Spain, in 2004. He has been a Lecture Professor with University of Zaragoza, since 2000. His current research interests include two main aspects on integration problems of network-based systems and the use of novel techniques and methodologies for solving them and the application of formal analysis techniques to the mining of event logs and databases.

JOAQUÍN EZPELETA received the M.S. degree in mathematics and the Ph.D. degree in computer science from the University of Zaragoza, Spain. He is currently a Professor with the Department of Computer Science and Systems Engineering, University of Zaragoza, where he conducts lectures on formal methods for sequential and concurrent programming and service-oriented architectures. His research interests include problems of modeling, analysis, and control synthesis for concurrent systems, the application of formal techniques to help in the development of correct distributed systems based on Internet and cloud technologies, and further the parallel processing of data and compute-intensive problems.