IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 3, 2017, accepted May 16, 2017, date of publication May 24, 2017, date of current version July 17, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2707600

Analysis of Users' Behavior in Structured e-Commerce Websites

SERGIO HERNÁNDEZ, PEDRO ÁLVAREZ, JAVIER FABRA, AND JOAQUÍN EZPELETA
Department of Computer Science and Systems Engineering, Aragón Institute of Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain
Corresponding author: Sergio Hernández (shernandez@unizar.es)

This work was supported by the Spanish Ministerio de Economía y Competitividad under Project TIN2014-56633-C3-2-R.

ABSTRACT Online shopping is becoming more and more common in our daily lives. Understanding users' interests and behavior is essential to adapt e-commerce websites to customers' requirements. The information about users' behavior is stored in the Web server logs. The analysis of such information has focused on applying data mining techniques, where a rather static characterization is used to model users' behavior, and the sequence of the actions performed by them is not usually considered. Therefore, incorporating a view of the process followed by users during a session can be of great interest to identify more complex behavioral patterns. To address this issue, this paper proposes a linear-temporal logic model checking approach for the analysis of structured e-commerce Web logs. By defining a common way of mapping log records according to the e-commerce structure, Web logs can be easily converted into event logs where the behavior of users is captured. Then, different predefined queries can be performed to identify different behavioral patterns that consider the different actions performed by a user during a session. Finally, the usefulness of the proposed approach has been studied by applying it to a real case study of a Spanish e-commerce website. The results have identified interesting findings that have made possible to propose some improvements in the website design with the aim of increasing its efficiency.

INDEX TERMS Data mining, e-commerce, web logs analysis, behavioral patterns, model checking.

I. INTRODUCTION

In today's ever connected world, the way people shop has changed. People are buying more and more over the Internet instead of going traditional shopping. E-commerce provides customers with the opportunity of browsing endless product catalogues, comparing prices, being continuously informed, creating wishlist and enjoying a better service based on their individual interests. This increasing electronic market is highly competitive, featuring the possibility for a customer to easily move from one e-commerce when their necessities are not satisfied [1], [2]. As a consequence, e-commerce business analysts require to know and understand consumers' behavior when those navigate through the website, as well as trying to identify the reasons that motivated them to purchase, or not, a product [3]–[5]. Getting this behavioral knowledge will allow e-commerce websites to deliver a more personalized service to customers, retaining customers [6] and increasing benefits [7].

However, discovering customer' behavior and the reasons that guide their buying process is a very complex task [3]. E-commerce websites provide customers with a wide variety of navigational options and actions: users can freely move through different product categories, follow multiple navigational paths to visit a specific product, or use different mechanisms to buy products, for example. Usually, these user activities are recorded in the web server logs [3], [8]. Web server logs store, in an ordered way, the sequence of web events generated by each user (commonly known as *click-streams*). The very valuable users' behavior is hidden in these logs, which must be discovered and analysed [9]. A correct analysis can be subsequently used to improve the website contents and structure [10], to adapt and personalize contents [11]–[13], to recommend products [14], [15], or to understand the interest of users in specific products [16], for instance.

Data mining techniques have proved their usefulness for discovering patterns in log files (when applied to the analysis of web server logs the term *web usage mining* [17] is used). Its main goal is to discover usage patterns trying to explain the users' interests. Different techniques have been successfully used in the field of e-commerce, such as classification techniques, clustering, association rules or

sequential patterns [18], [19]. In many application domains these techniques are used in conjunction with process mining techniques. Such techniques are part of the business intelligence domain and apply specific algorithms to discover hidden patterns and relationships in large data sets [20].

An e-commerce website is an open system where almost any customer behavior is possible. This flexibility makes the discovery of a process-oriented model representing customers' behavior a difficult task [21]. This is so because there are so many different possible interactions that the final process model is either an overfitting *spaghetti* model or an underfitting *flower* model [20], from which no useful analysis can be done. As a consequence, data mining techniques have been preferred for the analysis of e-commerce websites. Nevertheless, today's data mining techniques and tools have some constrains from the analysis point of view. On the one hand, they do not work in a direct way with the sequences of events (the click-stream and all the data associated to each click) generated during the user's navigation through the website, but with an abstraction of such sequence, a kind of global photograph that ignores causality relations. Such abstraction describes what happened during the session of a customer by means of a set of summarized data, such as the number of visited web pages, the frequency with which each product category was visited, or the time customers spend on a web page or category, for instance. On the other hand, most techniques are only able to classify these abstractions or discover simple relationships among certain high-level events of interest.

In this paper we propose the use of Temporal Logic and model checking techniques as an alternative to data mining techniques. Such techniques have proved their applicability for open systems [22]–[24]. We propose here a methodology for using it in structured e-commerce websites. The goal is to analyse the usage of e-commerce websites and to discover customers' complex behavioral patterns by means of checking temporal logic formulas describing such behaviors against the log model. At the beginning, web server logs are preprocessed to extract the detailed traces (sequences of events of a user session). Events can be user or system actions performed when a client visits a product or product category page, when he or she adds a product to the wishlist, when the search engine is used, etc. The business analyst can use a set of (predefined) temporal logic patterns to formulate queries that could help him to discover and understand the way clients use the website. Considering the website structure and contents as well as the different types of user's actions, these queries can check the existence of complex causality relationships between events contained in the client sessions. From the tool point of view, the necessity of having control on the way the checking algorithms are applied, as well as the disappointing performance results we obtained when using some model checking tools at our disposal, mainly when used against big models, drove us towards the interest of developing a specific model checking tool. We did it using the SPOT libraries for LTL model checking [25].

As a use case of the proposed approach we describe the analysis carried out for the Up&Scrap[1] e-commerce website, an important on-line Spanish provider of scraping products. The case of study describes the way raw logs have been processed, how the traces have been extracted, how users' behavioral patterns have been formulated and checked against the log. We also provide with some possible interpretations of the results obtained for the queries as well as some possible actions which could help in the re-design of the website whose aim is to improve it.

The remainder of this paper is organized as follows. In Section II, the related literature regarding techniques to analyse e-commerce web logs is reviewed. Section III introduces the concepts of linear temporal logic and model checking, and discuss its applicability to e-commerce logs. Up&Scrap, the enterprise used as case study in this work, is briefly presented in Section IV. Next, the methodological approach used to analyse the Up&Scrap logs is shown. Specifically, Section V shows the process followed to preprocess the web server logs and Section VI shows the different queries performed to analyse the logs and their interpretation. Finally, Section VII presents the paper's conclusions and makes suggestions for future research.

II. RELATED WORK

In the field of e-commerce, most data mining techniques process server logs to extract the sequences of user navigation events. Nevertheless, these sequences are not directly mined; instead, each sequence is transformed into a session characterization. A characterization usually consists of a set of high-level data summarizing what happened during the user's navigation. The contents of these structures can be diverse. In [10] the characterization contains the web browser used by the customer, the number of visited webpages, the time the customer spent on each page, or the keywords used in search engine; [14] and [16] focus on the users' interest in the different product categories and their characterization consist of the list of visited categories and the frequency of such visits. Unlike the previous approaches, [12] uses text mining techniques to discover the most frequent words contained in the Web pages a customer visits, generating the session characterization from these words. This solution tries to identify the user's interests from the contents of the visited pages. Another proposals build the characterization from questionnaires fulfilled by customers [26] or use a combination of customer's purchasing, demographic and personal data [15]. In any case, once customers' characterizations have been computed, clustering algorithms are generally used to discover the sets of sessions showing a similar behavior or some common interests. This information can subsequently be used to improve the website contents and structure [10], to adapt and personalize contents [12], [27], to recommend products [14], [15], to understand customers' behavior related to the buying process [26], [28], [29] or to

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

understand the interest of users in specific products [16], for instance.

Another researchers apply alternative mining techniques to predict the user's behavior. References [30] and [31] extract the users' navigational sequences to create statistical and probabilistic models able to predict the user next click. These models are represented as Markov chains. Nevertheless, these approaches present some drawbacks: the process of creating these models is computationally very expensive, and, besides, this type of models responds to very short-term reasoning (the model does not have information to know how the current navigational state has been reached and how future states representing long-term goals can be reached). The combination of clustering algorithms and Markov chains improves the predictions of these statistical models, as shown in [31]. The idea is to first group user sessions applying some clustering algorithms and, after, to generate a specific Markov chain for each of the obtained clusters.

As a complementary alternative to data mining techniques, process mining techniques try to also obtain causal relations between the events of users sessions. An example of such solution in the domain of e-commerce is presented in [21]. However, the open nature of the e-commerce websites, where the user can freely navigate through the different web pages, makes those techniques looking for a workflow-like model describing user behavior (such as a Petri net or a BPMN model) to be non adequate. As a consequence, the techniques used in [21] can only consider events with a very high abstraction level, having difficulties in the identification of patterns related to infrequent behaviors (as the case of buying events). As an alternative, a *declarative* approach can be adopted: instead of considering the system as following some imperative guidelines, a set of constraints is assumed so that anything that does not violates them is acceptable. The important question here is to discover such constraints [22]–[24]. These constraints are usually expressed by means of some temporal logic. In order to make the task of specifying properties easier for the analyst, a set of patterns, strongly related to some usual workflow structures, is used, as defined by the Declare property description language [32]. Declare facilitates the description of simple causality relations among the events in a user session. Alternative, MP-Declare [33] extends Declare by introducing the possibility of defining data and time constraints in the Declare patterns. However, it is limited to a predefined set of patterns and general formulas cannot be checked. Addition of new patterns would require the implementation, by the analysts, of instances of specific functions checking the desired behavior. Independently of the set of patterns used, the described temporal logic properties must be checked against the website log, which is carried out by means of some model checker tool.

Currently, there are powerful commercial tools for analysing logs of e-commerce websites, being Google Analytics [34] one of the main ones. Google Analytics controls the network traffic, collects information about user sessions (first and last web page visited, pages visited, time spent on each page, etc.), and displays reports synthesizing users' behavior. These traffic-based data can also be combined with other users' personal and geographic information. Google Analytics is not able to import the web server logs of a website, but it works analysing the information collected by means of page tagging techniques. These techniques have some disadvantages with respect to the log-based analysis, such as dependence on JavaScript and cookies, the necessity of adding page tags to every page, the complexity of tag-based implementations, the fact that, as a result, customers may experience a change in the download time of the website, or privacy concerns, for instance. Nevertheless, Google reports are rich in data that, in turn, require experts in the problem domain to exploit them. In any case, the conclusions of the analysis can be used to improve the website design, to design advertising and marketing campaigns, to analyse customers demographic information or to control real-time traffic. Similar commercial tools are *Clicky* [35], *Piwik* [36], *Adobe Analytics* [37] o *W3Counter Web Analytical tool* [38].

The methodology and tool proposed in this work try to overcome some of the drawbacks of the previous approaches, providing with the possibility of getting a very accurate interpretation of users' behavior:

- In comparison to the clustering approaches and the commented commercial tools, the advantage of our mining technique is that this provide causal relations among events of a user trace, instead of providing with a global view of the whole session. Besides, it is the fact of avoiding the need of tagging the web pages.
- With respect to those approaches whose main objective is predicting the coming possible events (as the case of Markov models, for instance), the approach allows having a global view of the sessions, making easier a global analysis of the user behavior, giving hints and facilitating the re-design of the website for a better adaptation to the user necessities.
- An interesting feature of the approach followed in this paper is that it properly fits the open nature of the use of e-commerce websites, where there are very few constrains for the users to navigate among site web pages.
- Another interesting feature of the followed mining approach is the fact of being able to analyse sequences of detailed events. The fact of considering the causal relations of events inside a user session, allowing to look for intra-session patterns (and not only patterns repeated in different sessions) can provide the analysts with a much more detailed perspective of a user behavior.
- The tool considers an event as a complex entity, seen as the conjunction of a set of attributes (usually configurable when specifying which information is incorporated to the log). This allows not only having a detailed view of the user activities, but also a (hierarchical) view with multiple aspects (it is a matter of proposing different LTL formulas involving the desired attributes in which we are interested).

Id	IP	Timestamp	Event name	Relative URL	Operation	Code	L1 section	L2 section
1	1.2.3.4	04/Mar/2016:03:36:50 +0100	Visit main section L1	/papeles	GET	200	papeles	
1	1.2.3.4	04/Mar/2016:03:36:58 +0100	Visit main section L2	/papeles/estampados	GET	200	papeles	estampados
1	1.2.3.4	04/Mar/2016:03:37:15 +0100	Add product to the cart	/checkout/cart/add/...	POST	200		
2	5.5.5.5	04/Mar/2016:03:36:59 +0100	Visit secondary section L2	colecciones/distress-crayons	GET	200	colecciones	distress-crayons
2	5.5.5.5	04/Mar/2016:03:37:17 +0100	Visit secondary section L1	/colecciones	GET	200	colecciones	

FIGURE 1. Extract from the final log used for the analysis of the Up&Scrap web server logs.

III. MODEL CHECKING TO ANALYZE EVENT LOGS

Let us now introduce linear temporal logics and model checking and briefly explain how it can be applied to the analysis of event logs in order to identify behavioral patterns paying special attention to the case of e-commerce web logs. Furthermore, in this section we will introduce some details of the model checker developed to enable this analysis.

A. BASICS ON LINEAR TEMPORAL LOGIC AND MODEL CHECKING

Usually, a program execution is viewed as the sequence of state transformations moving from a given initial state to a final state. We are considering a program state in terms of boolean formulas over a set of atomic propositions $\mathcal{A}$. Each atomic proposition is assumed to mean the truth of some property. The execution of a program sentence means that the values of some atomic propositions can change, moving from a boolean formula to another one. Therefore, talking about the program behavior requires to be able to talk about program states and also state evolutions.

Temporal logics have been used to talk about such evolutions [39]. Besides using propositional logic to talk about states ($(v_1 \wedge v_2) \vee \neg v_3$)), temporal logics add temporal (causality) operators, such as *next* ($\bigcirc$) and *until* ($\cup$). As a way of simplifying the writing of temporal logic formula, additional operators, such as *eventually* ($\Diamond$) and *always* ($\Box$), are used (which are defined in terms of the former ones).[2]

A program execution can be seen as the ordered sequence of the boolean formulas satisfied by the successive states the program reaches. This execution order is considered as the temporal structure. Having the finite set of possible program executions allows the analysis of the program behavior. *Model checking* techniques have been developed to carry out such analysis. These techniques check the truth of a set of behavioral *specifications*, stated in terms of temporal logic formulas, against the system *model*, which is composed of the set of possible executions [40], [41].

In this paper we are going to use Linear Temporal Logic, which defines a logic for (infinite) traces corresponding to program executions, with the approach followed in [42]. Let $\mathcal{A}$ be a given set of atomic propositions. The formal syntax of the set of correct LTL formulas is recursively defined as follows: 1) every $a \in \mathcal{A}$ is a LTL formula 2) if f and g are LTL formulas, then also $\neg f$, $f \vee g$, $f \wedge g$, $\bigcirc f$ and $f \cup g$ are.

[2]In the literature, X, G, F are used as alternative symbols for $\bigcirc, \Box, \Diamond$, respectively.

From a semantic point of view a LTL formula must be interpreted over runs of a program. A *finite state program* is a tuple $P_{\mathcal{A}} = (S, \rightarrow, s_0)$ where S is the finite set of program states, $s_0 \in S$ is the initial state and $\rightarrow \subseteq S \times 2^{\mathcal{A}} \times S$ is the transition relation, which describes the actions available at a given state and the state transitions corresponding to the execution of such actions. A *run* of $P_{\mathcal{A}}$ is an infinite sequence $\rho = s_0 \xrightarrow{x_0} s_1 \xrightarrow{x_1} s_2 \xrightarrow{x_2} s_3 \xrightarrow{x_3} \ldots$ where $(s_j, x_j, s_{j+1}) \in \rightarrow$ for any $j \geq 0$. Since we are interested in talking about log traces, for a run ρ, let us define $tr(\rho) = x_0 \cdot x_1 \cdot x_2 \ldots$ as its *trace*. Notice that the trace is an infinite word over the alphabet $2^{\mathcal{A}}$, of the possible subsets of $\mathcal{A}$. In the following, we will denote as σ_i the suffix of σ starting a i (notice that $\sigma = \sigma_0$).

Let f and g be two LTL formulas and $\sigma = x_0 \cdot x_1 \cdot x_2 \ldots$ be a trace. The satisfaction relation $\models$ is defined recursively as follows: 1) $\sigma \models p$ if $p \in \mathcal{A} \cap x_0$; also, $\sigma \models \mathit{true}$ and $\neg(\sigma \models \mathit{false})$; 2) $\sigma \models \neg f$ if $\neg(\sigma \models f)$; 3) $\sigma \models f \wedge g$ if $\sigma \models f$ and $\sigma \models g$; 4) $\sigma \models \bigcirc f$ if $\sigma_1 \models f$; 5) $\sigma \models f \cup g$ if there exists $i \geq 0$ such that $\sigma_i \models g$ and $\sigma_j \models f$ for any $j < i$. Stating that $\sigma \models f$ means that σ satisfies f.

B. APPLYING MODEL CHECKING TO EVENT LOG ANALYSIS

Let us now move from this conceptual framework to the case of event log analysis. A trace can be considered as the run of a program, where the set of atomic propositions corresponds to the set of events or event attributes [20], [32]. Let us focus on the portion of the event log shown in Figure 1, extracted from the case study analysed in this paper.

Each row corresponds to an event, where columns correspond to event attributes (the elements of a column can be considered as instances of the same attribute class). Events are ordered according to the time each one took place. We are considering the set of ordered events corresponding to the same session, those with the same Id, as a program run (*trace*). We associate an atomic proposition to each attribute value appearing in events, calling $\mathcal{A}$ to the whole set. The sequence of events of a trace can be seen as a sequence of elements belonging to $2^{\mathcal{A}}$.

In order to enable the use of model checking techniques, traces, which are finite, must be transformed into infinite ones. To achieve this, there are different proposals in the literature. The most commonly used is the addition of a final loop with a dummy *End* event to every terminal state [42]–[44]. Doing so, traces are now infinite and model checking can be applied. In fact, for each trace we have added a transition to

a dummy final state, as well as a self-loop for this state, both labelled with the *End* event and the conjunction of all the atomic variables, negated. The model checker must take into account that transformation in order to avoid interpretation mistakes [45].

C. MODEL CHECKER IMPLEMENTATION

In order to enable the application of LTL-based model checking on event logs, we have developed a log analysis system composed of two main components offered as REST Web Services. More details about the model checking analysis architecture can be found in [46]. First, the *Model Generator* uploads and transform the input log file, specified as a Comma Separated Values (CSV) file, so that it can feed the checker. Second, the *Model Checker*, which loads and analyses the previous file. The model checker has been implemented using the SPOT libraries for LTL model checking [25].

Besides usual temporal logic formulas, the tool provides with the possibility of defining sets of variables and macros to make easier the writing of LTL formulas [46]. Subsets of variables can be defined in multiple ways: by enumeration, as a range of identifiers or by means of regular expressions. Once a set VAR is defined, the appearance of ''?VAR'' in a formula means that the formula must be evaluated for each element belonging to ''VAR'', Thereby, as many formulas as elements in the set VAR are automatically checked by the tool. Similarly, macros can be defined on these sets as a logical OR or AND between all the elements on the set. For example the macro ''_OR or_var VAR'' indicates that the appearance of ''?or_var'' is replaced in the formula by the logical ''OR'' of all the elements in the set ''VAR''. Also, some formulas can be defined with a given name, avoiding the same formula to have to be written more than once.

D. APPLYING MODEL CHECKING TO THE ANALYSIS OF E-COMMERCE WEBSITES

Users of any e-commerce site navigate through the different web pages executing two types of interactions: either a GET operation to retrieve some information or a POST operation, usually requesting the website to execute some action, such as adding some product to the cart, buying some product, logging in, etc. The website log records such actions together with some associated information, such as the IP the user is connected from or the time at which the interaction occur [47], for instance. Some of these actions correspond to events that are common to any e-commerce website such as the ones related to visiting the sections containing products. Therefore, a general way of classifying the events in the web logs according to the product categorization can be proposed. From now on, we are going to describe the proposed approach to relate the website structure and the events in the log, to identify meaningful set of events, and to ask for behavioral usage patterns using model checking based on the previous classification.

To apply model checking techniques, we are going to associate temporal logic formulas to events, which will allow us to see the log as a Kripke structure representing the model to be analysed [22], [32]. For that, we are first going to define the set of atomic propositions, and transform events into conjunctions of such variables. This will be done during the pre-processing phase (as described in Section V), whose output will be the model representing the log.

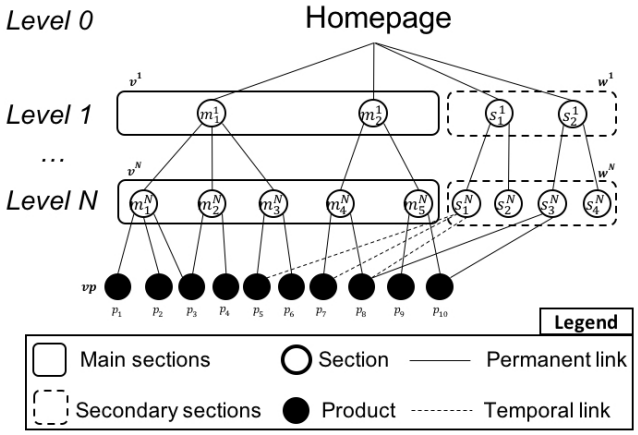

FIGURE 2. Typical structure used to categorize products in an e-commerce website.

Figure 2 shows the typical structure used in e-commerce websites to organize and categorize products. Similar taxonomies have been proposed by different authors but including only main sections [7], [16]. From the homepage (level 0) different sections can be accessed (level 1). Two different types of sections can be distinguished.

1) *Main sections*, which correspond to the main product categorization. These sections allow access to all products. In general, the product categorization is usually disjoint, but this is not mandatory: in some e-commerce websites the same product could belong to different sections.
2) *Secondary sections*, which provide a secondary categorization of the website products, whose objective is to allow the access to a subset of products with some common and specific features. Unlike the previous case, not all products must be accessible from these secondary sections. Furthermore, we can distinguish two different types of secondary sections depending on whether products in such sections are permanently or temporarily added to the section. An example of sections with temporary products would be offers or sections with new products that are periodically renewed. An example of secondary sections with permanent links would be sections where you can access products by manufacturer, theme, etc.

Independently of its type, each section is usually split into several subsections to refine product classification. Each e-commerce website establishes its own organization (categories, levels, etc.).

The previous structure is reflected in the website navigational map: each section corresponds to a specific web page

from where products and other sections can be accessed. Regarding the navigational structure, the links shown in Figure 2 are not the only way to navigate through the website since many links and menu options are usually provided to facilitate and improve the user browsing experience. Thereby, products could also be directly accessed from all levels, not only from the last one, some levels can be skipped and horizontal links between sections could also be included. Alternatively, some products could also be accessed by using other standard mechanisms included in e-commerce websites such as search engines. These mechanisms are analogous to secondary sections since they share the same goal of providing and alternative way of visiting products. As a consequence, in the proposed classification they are considered as another secondary section.

Since each section corresponds to a different web page, each section has a unique URL that enables to identify the visiting event. The website designer knows the relation between the URL and the associated sections in the structure shown in Figure 2. In many cases this relation is clearly reflected in the URL itself. For instance, in the case we are studying in this paper, the second event in Figure 1 is related to visiting relative URL */papeles/estampados*, belonging to the main level-2 category *estampados* which is a subcategory of main level-1 category *papeles*. When not, the system manager must establish the mapping between web-pages and atomic propositions.

Similarly, for POST requests the same approach can be used by assigning to each POST request a specific event type. Section V describes in a more detailed way the pre-processing phase.

Let us now introduce some notations identifying the main elements in the website and correlate them with the web server logs.

Let $N \in \mathbb{N}$ be the number of levels in the structure. Let $\overline{V} = \{v^i \mid i \in 1..N\}$ ($\overline{W} = \{w^i \mid i \in 1..N\}$) be a set of atomic propositions, bijective with the set of main (secondary) sections of the different levels, used to identify event types associated with visiting a main (secondary) section. Each event corresponding to a given main (secondary) level-i section will be annotated with the v^i (w^i) proposition.

For each main section, let M^i (S^i) be a set of atomic propositions bijective with the main (secondary) sections of level i. According to that, the event corresponding to visiting m^j main section of level $j \leq N$ can be represented as the logic formula $v^j \wedge m^1 \wedge m^2 \wedge \ldots \wedge m^j$, with $m^\alpha \in M^\alpha$ for each $\alpha \in 1..j$. For the case of secondary sections, the same approach will be adopted, using $\overline{W}$ and S^i sets.

Regarding products, all the events related to downloading a product web-page are described using the same atomic proposition *vp* (view product). Alternatively, different propositions could be used to distinguish products. However, doing so, there would be too many different entities, making the analysis not only very difficult or even impossible, but also the obtained results less interesting than when working with categories associated with the different sections.

In order to make easier referring to different events, let us now introduce some additional notational conventions:

- Let $M = \bigcup_{i=1}^{N} M^i$ ($S = \bigcup_{i=1}^{N} S^i$) be the set of main (secondary) sections.
- For any $i \in 1..N$, let $V^i = \{v^i \wedge m^1 \wedge m^2 \wedge \ldots \wedge m^i \mid m^\alpha \in M^\alpha \text{ for any } \alpha \in 1..i\}$ be the set of events corresponding to visiting level-i main sections with the information of the product categorization.
- For any $i = 1..N$, let $W^i = \{w^i \wedge s^1 \wedge s^2 \wedge \ldots \wedge s^i \mid s^\alpha \in S^\alpha \text{ for any } \alpha \in 1..i\}$ be the set of events corresponding to visiting level-i secondary sections with the information of the product categorization.
- Let *vh* be the atomic proposition associated to visiting the homepage.
- Let *vp* be the atomic proposition associated to visiting products.
- Let A be the set of events corresponding to POST actions such as adding products to the cart, adding products to the wishlist, logging in, buying products, etc.

Defining different common events and sets of events enable the possibility of proposing pattern queries that can be explored independently of the e-commerce website analysed. As stated above, model checking can be used to query about specific states, their evolution and their relations. In the context of e-commerce web server logs this means that model checking can be used to analyse the sections visited by users, the navigational paths followed when accessing specific pages of the website, the relation between different web sections or the sections that lead users to buy products, for instance. Based on the defined structure and sets, we are going to define three different types of query patterns.

1) **Queries related to the analysis of the web sections visited by users.** This first type comprises the most simple queries: which sections do the users visit? Examples of interesting queries in this regard could be:
   - Sessions where the user visits main level-1 sections. For any $v \in V^1$, the LTL formula $\Diamond(v)$ will give the number of traces in which the user visits category e of the main sections of level 1.
   - Sessions where the user visits some of the secondary sections: $\Diamond(w^1 \vee w^2 \vee \ldots \vee w^N)$ will count how many sessions contain at least one event with one of the secondary atomic propositions.
   - Sessions where users exclusively visit level-1 main sections (no secondary section is visited): $\Diamond(v^1) \wedge \neg\Diamond(w^1 \vee w^2 \vee \ldots \vee w^N)$
2) **Queries identifying navigational patterns.** These queries try to establish causal relations between the fact of visiting different sections in the website:
   - Specific sections visited right after visiting a level-1 main section: for any $e \in (V \cup W)$, $\Diamond(v^1 \wedge \bigcirc e)$ counts the number of traces in which the next event after visiting main level-1 section contains e.

- Level-2 main sections that are not visited before the corresponding level-1 main section is visited: for any $m \in M^1$, $\Diamond(v^2 \wedge m) \wedge (\neg(v^2 \wedge m)\ U\ (v^1 \wedge m))$ gives the number of traces which visit $v^2 \wedge m$ (section m of the main level-2 categories), which also visit $v^1 \wedge m$ (section m of the main level-1 categories), but $v^2 \wedge m$ only happens later than $v^1 \wedge m$.

3) **Queries to correlate user navigational patterns with user actions.** These queries try to correlate specific actions, different than visiting sections, with the behavioral patterns previously identified. Example of these queries are the following ones:

- Sessions that perform a specific action without visiting main sections: for any $a \in A$, $\neg\Diamond(v^1 \vee v^2 \vee \ldots \vee v^N) \wedge \Diamond(a)$ counts the number of sessions where the action a is performed without visiting any main section.
- Sessions where a specific action happens right after visiting a secondary section: for any $e \in W$, for any $a \in A$, $\Diamond(e \wedge \bigcirc a)$ counts how many sessions perform the action a right after e happened.

For the analysis of e-commerce websites by using web logs there are two interesting issues. First, identifying the most interesting website sections and the relation between them [10], [16]. Second, analysing the navigational paths followed by users to look for products [12]. This information can be used with different purposes such as improving the website contents and structure, to generate recommendations for users or to predict next clicks. By using model checking techniques and applying the different type of queries we have identified, the website sections that are more interesting for users can be detected. Different navigational and usage patterns can be discovered to analyse how users are behaving while accessing the website and how they are interacting with the most relevant web sections. Finally, the findings can be correlated with interesting actions involved in the use of the e-commerce website such as adding products to the cart or buying them.

IV. UP&SCRAP

Let us now introduce the company that motivates this work and the main structure of its website. Up&Scrap is the leader company in Spain for the sale of equipment for scrapbooking. It was founded in 2012, it has more than 25,000 clients and has performed over 85,000 shipments. The company's website[3] includes over 2,500 products that can be accessed and purchased from many sections. Figure 3 shows a screenshot of the web homepage. The structure of the website is as follows:

- *Main sections*. The website includes a main menu with the main sections. From this menu both, categories and subcategories, can be accessed. There are 8 sections (they are listed in the same order as they appear in Figure 3):
  - *Papers*. This category provides lots of printed papers to perform scrapbooking projects. It includes 2 subcategories.
  - *Decorate*. This category offers a wide range of ornaments for scrapbooking projects, such as wooden and metal ornaments, washi tape, buttons, ribbons, pearls, thickers, etc. This category includes 16 subcategories.
  - *Stamp*. In this category all the products required to stamp and paint are offered, including a complete range of inks and different sets of stamps. This category includes 8 subcategories.
  - *Tools*. This category provides any tool required for scrapbooking like scissors, glue, cutter, tape, etc. It includes 9 subcategories.
  - *Project life-smash*. This category provides products to create your own albums with photographs and memories of your life creatively. It includes 2 subcategories.
  - *Albums*. The section includes albums and covers of different sizes, as well as tools to create you own albums. It includes 3 subcategories.
  - *Home decor-diy*. This category offers products to decorate your home in a personalized and different manner. It includes 11 subcategories.
  - *Gifts*. This category includes different starter kits, tool kits and gift vouchers. It does not include subcategories.
- *Secondary sections*. They are composed of sections to show new products and special offers, as well as secondary categorizations according to multiple criteria. We can identify 6 sections.
  - Sections with temporary products.
    - *Offers*. This sections allows users to directly visit offers. The section includes 7 subcategories with the same classification used in the website main sections (except gifts).

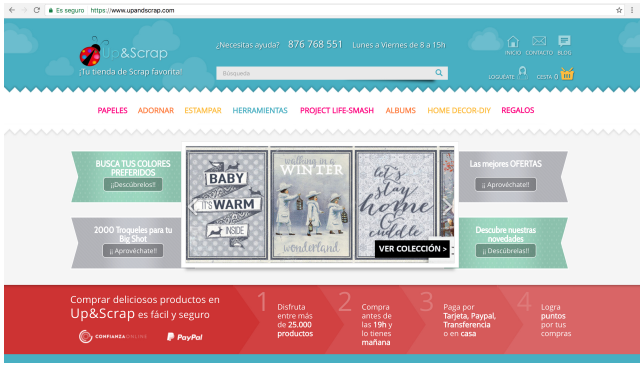

FIGURE 3. Homepage of the Up&Scrap website.

[3]http://www.UpAndScrap.com

```
1.2.3.4 - - [04/Mar/2016:03:36:50 +0100] "GET /papeles HTTP/1.1" 200 28883 "https://www.upandscrap.com" "Mozilla/5.0 (Linux; Android 5.1.1; SM-G531F Build/LMY48B) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.95 Mobile Safari/537.36" "-" "0.763"
1.2.3.4 - - [04/Mar/2016:03:36:58 +0100] "GET /papeles/estampados HTTP/1.1" 200 28945 "https://www.upandscrap.com/papeles" "Mozilla/5.0 (Linux; Android 5.1.1; SM-G531F Build/LMY48B) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.95 Mobile Safari/537.36" "-" "0.808"
5.5.5.5 - - [04/Mar/2016:03:36:59 +0100] "GET /colecciones/distress-crayons HTTP/1.1" 200 19231 "https://www.upandscrap.com/colecciones" "Mozilla/5.0 (iPhone; CPU ,,iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75 Safari/601.1" "-" "0.743"
1.2.3.4 - - [04/Mar/2016:03:37:15 +0100] "POST /checkout/cart/add/uenc/aHR0cHM6Ly9BhZG9zL3BhZ2luYS8xNS8,/product/26731/form_key/b97404f6593eb808/ HTTP/1.1" 200 128 "https://www.upandscrap.com/papeles/estampados" "Mozilla/5.0 (Linux; Android 5.1.1; SM-G531F Build/LMY48B) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.95 Mobile Safari/537.36" "-" "0.728"
5.5.5.5 - - [04/Mar/2016:03:37:17 +0100] "GET /colecciones HTTP/1.1" 200 28534 "https://www.upandscrap.com/colecciones/distress-crayons" "Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75 Safari/601.1" "-" "0.796"
```

FIGURE 4. Extract from the raw log of the web server, where IP addresses have been anonymized.

  - ∗ *New products*. The new products, recently added to the e-commerce catalogue, are listed in this section. The section includes 7 subcategories with the same classification used in the website main sections (except gifts).
- Sections with permanent products.
  - ∗ *Brands*. This section provides a list of all the product brands. Users can select a brand to view its products. Currently, there are 151 brands.
  - ∗ *Thematics*. This section classifies products according to different and diverse thematics such as summer, babies, dogs, etc. Currently, there are 58 different thematics.
  - ∗ *Collections*. This section groups products according to collections. Currently there are 543 collections.
  - ∗ *Designers*. This section allows users to visit the products of their favourite designers. Currently, the products of 9 designers can be accessed from this section.

- Additionally, the website includes a *search engine* that allows users to directly look for products without using the proposed sections.

The website, despite its structured design, allows users to reach products following many different navigation paths or, even, directly from the search engine. Analysing users' behavior would be of a great help to improve users' interactions with the website.

In the following sections we are going to describe the methodology used for the analysis of the Up&Scrap website logs. The approach starts retrieving the logs from the server and preprocessing them to obtain a log ready for the model checking analysis. Next, we describe the process used and the queries performed with the goal of detecting behavioral patterns and identifying user interests. Finally, we discuss the results obtained and how they can be used to improve the company's website.

V. DATA PREPROCESSING

The initial step of web usage mining analysis is data preprocessing [48]–[50]. The raw data have relatively low business value unless they can be transformed and processed to produce actionable knowledge [51]. Therefore, in order to enable the analysis, raw logs must be preprocessed to discard uninteresting requests, to identify user sessions and to prepare the log to enable its analysis.

The web logs follow the Common Language Format standard (CLF) [47] and provide raw information such as the IP address from which the session was established, the date and time of the request, the page URL or the HTTP status returned to the client, for instance. Figure 4 shows a piece of the raw web log of the considered use case, corresponding to two months of usage of the UpAndScrap website, containing 8, 607, 625 events.

The preprocessing step can be split into three main phases. The first two are common to any web usage mining project [10], [12]. The third one is introduced to prepare the log contents for applying the used model checking techniques. Let us describe that phases in more detail.

A. LOG CLEANING

The objective of this phase is to remove undesired records that may distort the results of the analysis. For that, the following steps are carried out:

- Removing automatic requests such as the ones performed by robots, spiders and crawlers. To do that, the IPs requesting the `robots.txt` file and the requests with a user agent belonging to automatic requests are deleted. Furthermore, requests corresponding to IPs demanding for a large number of pages in a short period of time are also removed since it can be assumed that they are performed by automatic tools.
- Deleting requests with erroneous status codes (4*xx* and 5*xx* codes). Since we are interested in navigational patterns, erroneous requests are not interesting in this regard.
- Discarding requests of irrelevant HTTP methods. Only GET and POST requests have been considered since they are the unique directly requested by users.
- Deleting requests asking for multimedia contents, since these requests are automatically requested by the browser.

After this first phase the log has been reduced to 5, 875, 479 records, a 68.26% of the original size.

B. USER IDENTIFICATION AND SESSIONIZATION

The aim of this phase is to group the events belonging to the same session (in terms of process mining, we are establishing

the traces of processes). For that, we have used a heuristic where a user session is composed of those events generated from the same IP address in a period of 25.5 minutes. This concrete value has been typically used in web usage mining approaches [49] and it is consistent with the behavior we have observed in the log. As a result of this phase, 144, 330 user sessions have been identified.

C. LOG PREPARATION

The aim is to prepare the log file to feed the model checker. For that two types of actions are performed. On the one hand, in the *categorization* sub-phase each record is analysed to identify high-level events and to extract meaningful information. On the other hand, in the *simplification* sub-phase, log contents are reduced to increase the effectiveness of the model checking techniques.

Regarding the log categorization, different events can be identified by analysing the CLF log contents. For such purpose, each event is automatically classified by considering whether it is a GET or a POST request and by analysing its URL in terms of the presence and/or absence of specific keywords and resources (that is, the words between slash characters). For the classification, we have used the strategy depicted in Section III where GET requests related to accessing some of the sections are classified depending 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 ($N = 2$).

For the remaining requests, we have identified different events based on their unique URL. We have identified 63 different types of events such as *Visit main section L1, Visit secondary section L2, Visit product, Login, Logout, Add product to the wishlist, Add product to the cart*, etc. These events refer to different actions and can affect to different sections of the website. However, not all the event types are interesting for the analysis since some of them provide superfluous information. They can, for instance, refer to user account management or to visiting the legal warning, for instance. Thus, in the following phase 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 carried on are considered.

Furthermore, different information is recorded in the final log along with the *event name*: the *IP* that makes the request, the *timestamp*, the *URL* requested (we only include the content after the first "/" since the previous content corresponds to the website address and it is common for all requests), the type of operation (GET or POST), the HTTP *status code*, the first level section *(L1 section)* and the second level section *(L2 section)* (if any). The corresponding section, in the case of events such as *Visit secondary section L2*, is automatically obtained since this information is coded in the URL as the first and second substring (using the character "/" as a separator), respectively.

Other interesting issue is that we cannot identify the category and subcategory to which a product belongs by using the web server logs. This is because in the Up&Scrap web page, the URL of a product only contains the name of the product as resource and does not provide any information about how the product has been categorized. To automatically obtain such information, web structure mining techniques could be explored [8], [19].

Despite the fact that the log can be analysed after the categorization phase, an additional *simplification* phase is performed. This phase has the goal of reducing the amount of information included in the log by filtering the records that do not contain relevant information. With that purpose three actions are performed.

First, sessions with less than three requests are discarded since 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 since they do not provide valuable information for the analysis. Since the goal of analysing the logs is to extract information about users' behavior and preferences when buying products, there are many events that can be considered as superfluous, such as events related to the user account management or rating the products. In this case, we have identified a set of 12 types of events that we consider relevant for the analysis and we have filtered the remaining ones. In the following we detail them according to the different sets identified in Section III.

- v^1 = Visit_main_section_L1, M^1 = {albums, decorate, gifts, home_decor–diy, papers, project_life-smash, stamp, tools}.
- w^1 = Visit_secondary_section_L1, S^1 = {brands, designers, collections, new_products, offers, search results, thematics}. Accessing the results of the search engine is considered as visiting a secondary section, as stated above.
- v^2 = Visit_main_section_L2, M^2 contains 51 sections, which we do not enumerate for the sake of simplicity.
- w^2 = Visit_secondary_section_L2, S^2 is composed of 779 sections (not enumerated here).
- vh = Visit_homepage is the event type representing that the homepage is visited.
- vp = Visit_product is the event type representing that the URL of a product is visited.
- A = {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, update_product_from_the_cart}.

Then, as a result, the model will have 857 atomic propositions. According to the previous definitions, an example of LTL formula that could be used to refer to the second event shown in Figure 5 is: "Visit_main_section_L2 $\wedge$ papeles $\wedge$ estampados" being "Visit_main_section_L2" the event type corresponding to visiting a level-2 main section, "papeles" the name of the corresponding level-1 main section and "estampados" the name of the specific level-2 main section visited.

Id	IP	Timestamp	Event name	Relative URL	Operation	Code	L1 section	L2 section
1	1.2.3.4	04/Mar/2016:03:36:50 +0100	Visit main section L1	/papeles	GET	200	papeles	
1	1.2.3.4	04/Mar/2016:03:36:58 +0100	Visit main section L2	/papeles/estampados	GET	200	papeles	estampados
1	1.2.3.4	04/Mar/2016:03:37:15 +0100	Add product to the cart	/checkout/cart/add/...	POST	200		
2	5.5.5.5	04/Mar/2016:03:36:59 +0100	Visit secondary section L2	colecciones/distress-crayons	GET	200	colecciones	distress-crayons
2	5.5.5.5	04/Mar/2016:03:37:17 +0100	Visit secondary section L1	/colecciones	GET	200	colecciones	

FIGURE 5. Extract from the log generated after the preprocessing phase.

As the last step we have deleted *duplicated* consecutive records, since they do not provide useful information, keeping with only one event instance. We have identified three situations were events can appear in a duplicated consecutive way:

1) The user reloads the web page or repeats the same click.
2) Events corresponding to visiting different pages of a given listing. When users are looking for products within a category, subcategory or search, new pages are automatically requested by simply scrolling down the product list. From a conceptual point of view, the user is repeating the same action, *looking at the products of a list*. We abstract this sequence as a unique event.
3) Duplicated events appearing after removing superfluous events. This is the case, for instance, when the user enters in the homepage to login in the system. Afterwards, the homepage is reloaded showing that the user is connected. In this situation, the event of visiting the homepage is duplicated because of filtering the events related with the login process. Therefore, the second event appears after removing the intermediate events, and can be discarded.

TABLE 1. Table summarizing the amount of records included in the log after the different preprocessing stages.

Log	Number of records
Raw log	$8,607,625$
Clean log	$5,875,479$
Filtered log	$3,680,882$
Final log (without duplicates)	$1,331,697$

Regarding the log size, discarding short sessions and events that are not relevant for the analysis reduces the number of records in the log to $3,680,882$ records. Finally, removing duplicated events reduces the log size to $1,331,697$ records, a 31.93% of the size of the previous log. This reduction is because most of the requests are due to the search of products in listings. Figure 5 shows a piece of the final log used as input for the model checker. Table 1 summarizes the amount of records in the log after the different phases of the preprocessing step. With respect to the original raw log, the size of the final has been reduced to a 13.66 % of the original size.

VI. IDENTIFYING USERS' BEHAVIORAL PATTERNS

Next we are going to detail the process carried out to analyse and identify behavioral patterns from the Up&Scrap logs. Prior to the analysis we have defined a set of variables and macros based on the sets identified in Section III-D. They are enumerated below, according to their equivalence with the sets proposed in Section III-D.

- *Variables*. Let `?M1` denote M^1, `?M2` denote M^2, `?M` denote $M^1 \cup M^2$, `?S1` denote S^1, `?S2` denote S^2, `?S` denote $S^1 \cup S^2$, `?V1` denote V^1, `?V2` denote V^2, `?V` denote $V^1 \cup V^2$, `?V_BAR` denote $\overline{V}$, `?W1` denote W^1, `?W2` denote W^2, `?W` denote $W^1 \cup W^2$, `?W_BAR` denote $\overline{W}$, `?VW1` denote $V^1 \cup W^1$, `?VW2` denote $V^2 \cup W^2$, `?VW` denote $V \cup W$ and `?VW_BAR` denote $\overline{V} \cup \overline{W}$.
- *Macros*. Let `?OR_V_BAR` a macro defined as $(\bigvee_{v \in V} v)$, `?OR_W_BAR` a macro defined as $(\bigvee_{w \in W} w)$, and `?OR_VW` a macro defined as $(\bigvee_{v \in (V \cup W)} v)$.

TABLE 2. Percentage of appearances of each event in comparison to the total number of events and the number of sessions in the log. Events in the table are shown lexicographically ordered.

Event	% of total number of events	% appearances in sessions
Add all wishlist products to the cart	0.01	0.04
Add product from the wishlist to the cart	0.03	0.17
Add product to the cart	7.27	14.19
Add product to the wishlist	1.07	2.26
Buy products in the cart	0.72	4.03
Delete product from the cart	3.10	6.17
Visit main section L1	8.14	22.27
decorate	1.08	3.99
albums	0.51	2.39
stamp	0.76	3.24
tools	1.52	6.22
home decor-diy	0.41	1.86
papers	3.15	9.69
project life-smash	0.38	1.75
gifts	0.31	1.89
Visit main section L2	27.22	48.05
Visit product	25.62	55.31
Visit root page	7.64	39.18
Visit secondary section L1	11.62	35.98
collections	0.20	0.52
designers	0.01	0.05
brands	0.09	0.34
new products	3.44	12.77
offers	1.37	5.63
search results	6.38	22.22
thematics	0.09	0.31
Visit secondary section L2	7.56	29.78

First of all we have performed a statistical analysis in order to identify the percentage of occurrences of each event and we have checked whether the same tendency is shown with respect to their appearance in different sessions. Table 2 shows the results of this analysis. The two main events are

Visit main section L2 and *Visiting product*. Users mainly use some of the subcategories from the main sections to navigate through the website contents. If we look at the number of sessions containing such events, around half of them visit some main subsection and the information of a product. The second preferred option to access the web contents is *the use of the search engine* with 22.22% sessions containing such event. Next, *visiting a level-1 main section* without filtering products by subsection is the most used approach.

In the following we are going to discuss the most relevant findings of the model checking-based analysis. The goal of this analysis is to identify meaningful usage patterns that could be used to improve the Up&Scrap website design. In this respect, we are mainly interested in detecting the most relevant and used parts of the website and the relationship existing between them, as well as identifying behavioral patterns related to the buying process. For that, we present the queries along with the question that they intend to answer, the results of the queries, the interpretation of the results and the actions proposed to improve the website, if any.

A. USAGE PATTERNS RELATED TO MAIN SECTIONS

As shown in Table 2, main sections are the main resources used to navigate through the website. That means that main sections are preferred to secondary ones. Let us analyse if users prefer to access to level-1 or level-2 sections.

Query 1 *What are the most visited main sections when taking into account both level-1 and level-2?*
The query "$\Diamond$(?OR_V_BAR $\wedge$?M1)" counts, for each main section $m \in M^1$, the number of sessions where m and any element of $\overline{V}$ is visited.
Are level-1 main sections preferred to level-2 main sections?
The query "$\Diamond$(Visit_main_section_L1 $\wedge$?M1)" counts for each main section $m \in M^1$ the number of sessions visiting it, while "$\Diamond$(Visit_main_section_L2 $\wedge$?M1)" counts, for each $m \in M^1$, the number of sessions visiting its subsections.

TABLE 3. Number of sessions visiting each main section, percentages of them directly visiting level-1 section and percentage of them directly visiting the corresponding subsections.

Main section	Number of sessions	% in L1	% in L2
Decorate	15,774	36.42	79.40
Albums	5,241	65.67	47.61
Stamp	21,874	21.29	92.15
Tools	34,206	26.16	91.39
Home decor-diy	10,059	26.59	84.52
Papers	25,176	55.35	67.41
Project life-smash	7,132	35.21	76.04
Gifts	2,720	100.00	-

Results Table 3 summarizes the results of that queries. The second column shows the total number of sessions visiting each section (first query results). The third column shows the percentage of the sessions visiting the level-1 section (second query results). The fourth column indicates the percentage of these sessions which visit some of its level-2 sections (third query results). Note that some sessions could visit both levels.

Interpretation Results show that *tools* is the most visited section followed by *papers* and *stamp*. If we focus on the individual accesses to the first or second level two different patterns can be observed. On the one hand, *albums* and *papers* show a similar percentage of accesses to both levels. On the other hand, *decorate*, *stamp*, *tools*, *home decor-diy* and *project life-smash* present a much higher percentage of accesses to the second level. In general, sections with a large number of subsections are accessed directly from the second level since there are too many different products inside the global sections. On the other hand, sections composed of a few subsections are visited directly. The exception is the section *project life-smash* that only contains two subcategories (*project life* and *smash*). This anomaly could point towards both subsections being not strongly related with each other and they could be considered as two different level-1 main sections.

Proposed actions It should be further studied if the categorization related to the section *project life-smash* is adequate since results suggest that the subsections of this category are not related. A solution could be to consider each level-2 section as a level-1 one.

Since most of sessions are accessing to level-2 sections, we are going to focus on access patterns related with such sections. Let us analyse the distribution of the accesses to these sections.

Query 2 *Is the access to level-2 main sections homogeneous?*
The query "$\Diamond$?V2" counts, for each $m \in M^1$ and each $m' \in M^2$ the number of visits to each level-2 section (remember the form of events in V^2 described in Section III-D).

Results Figure 6 shows the scatter-plot that summarizes the results, grouped by the corresponding level-1 main section (m). Data is presented considering for each level-2 section m' the percentage of sessions that visit that subsection with respect to the total number of sessions visiting some of the subsections of m.

Interpretation As it can be seen there are some level-2 subsections that concentrate the interest of users. This happens in all the categories except *decorate*, where the access is rather homogeneous. Furthermore, it is specially remarkable in sections with few subsections (*albums, papers and project life-smash)*. This leads to two different issues. First, some level-2 sections get very little attention. Probably they are too specific or just contain a few number of products. Second, level-2 sections with a large number of visits may contain too many products making it difficult for users to find the products the are interested on. Splitting such sections could be beneficial for that cases.

Proposed actions Modifying the product categorization should be studied to homogenize the accesses.

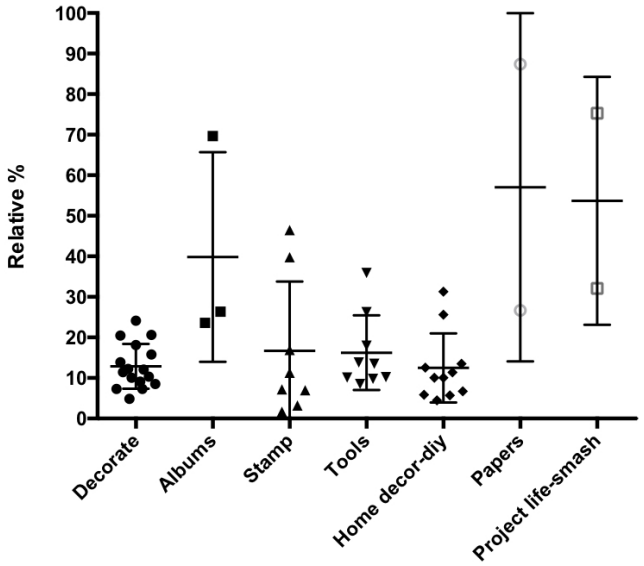

FIGURE 6. Distribution of the number of sessions visiting level-2 main sections grouping the values by their corresponding level-1 main section.

Two opposite actions could be carried out. On the one hand, sections with a low number of accesses could be grouped into more general ones. On the other hand, splitting the sections that focus the interest could make it easier for users to find products that are interesting form them.

B. USAGE AND NAVIGATIONAL PATTERNS

Next, we are going to explore some navigational patterns that illustrate users' preferences when browsing the website. Specifically, we analyse if sections are visited as the first option and if they are visited in an exclusive way.

Query 3 *What main or secondary sections are preferred by users as the first step to look for products?*
The query "$(\neg$?OR_VW_BAR$) \cup ($?OR_V_BAR $\wedge$?M1)" identifies, for each $m \in M^1$, the sessions where the own m or any of its subcategories is visited, while no other section has been visited previously. For secondary sections an analogous query is used replacing M1 by S1 and OR_V_BAR by OR_W_BAR.

Results Results are depicted in Table 4. As in the previous case, the total and the relative percentages for each event are presented.

Interpretation Results show that *collections* is the web section preferred by users and most of the accesses to this section are the users' first choice. This is specially remarkable since it is a secondary section. This behavior could be explained due to marketing and mailing campaigns that focus on promoting collections. Regarding main sections, the access pattern is consistent with the total number of sessions visiting each section. The results also show that the *gifts* section has a low interest for users as it only contains gift cards and starter kits. As for the rest of secondary sections, both the *search engine* and checking *new products* are common choices between users. Furthermore, most of the accesses to these sections are the users' first choice. This could show that they are accessed by regular users to look for new products. It is also important to note that the *offers* section is not attracting as much attention as it could be expected, since it is intended to attract new customers with cheaper prices as well as regular users that already know the product catalogue. Another interesting fact is that visiting a specific *brand* is also an explored option. This may indicate a set of users that are fans of specific brands and check their products regularly. Finally, accessing to *designers* as a first choice is not a common alternative. This is a sign of its low relevance to users.

TABLE 4. Sections that are selected as the first choice to search products. The results include both levels of product categorization.

	Section	Total %	Relative %
Main	Decorate	4.17	38.04
	Albums	1.41	38.60
	Stamp	7.38	48.49
	Tools	13.4	56.34
	Home decor-diy	3.18	45.46
	Papers	9.47	54.11
	Project life-smash	1.98	39.90
	Gifts	0.26	13.90
Secondary	Collections	16.02	88.45
	Designers	0.41	34.85
	Brands	5.43	57.90
	New products	9.81	76.83
	Offers	3.59	63.70
	Search results	10.18	45.82
	Thematics	1.55	46.80

Proposed actions The reason behind the large number of accesses as first choice to *collections* should be carefully analysed. The effectiveness and the interest of using this secondary section as the entry point for the website should be studied. The importance of the *Offers* section should be promoted within the website to attract both new and regular users. Marketing, advertising and mailing campaigns should be reviewed considering this information.

Next, we check which sections are visited in isolation, that is, sections that are the only section visited during a session.

Query 4 *Are sections visited exclusively?*
The query "$\Diamond$(?OR_V_BAR $\wedge$?M1) $\wedge \Box$(?OR_VW_BAR $\rightarrow$?M1)" is executed to answer the previous question. For each $m \in M^1$, the first part of the query checks that m or any of its subsections is eventually visited, while the second part imposes that always that a section is visited, the visited section is m (since $m \wedge m'$, being $m' \in M^1$, is not possible). As in the previous case, an analogous query is executed to analyse secondary sections replacing M1 by S1 and OR_V_BAR by OR_W_BAR.

Results Table 5 shows the results of the query. Total and relative percentages for each section are presented.

TABLE 5. Analysis of sections that are accessed in isolation.

	Section	Total %	Relative %
Main	Decorate	2.11	19.22
	Albums	0.89	24.33
	Stamp	4.11	27.02
	Tools	8.38	35.25
	Home decor-diy	2.23	31.94
	Papers	5.17	29.54
	Project life-smash	1.09	22.04
	Gifts	0.07	3.57
Secondary	Collections	13.79	76.14
	Designers	0.23	19.30
	Brands	3.26	34.73
	New products	5.48	42.92
	Offers	1.3	23.10
	Search results	3.59	16.14
	Thematics	0.81	24.52

Interpretation The results are similar to the ones representing accesses as the first choice (Table 4). However, they show two remarkable findings. First, a large portion of the accesses to *collections* is performed in isolation. Users that directly access a specific collection are not interested in another sections. Therefore, despite the fact of being a usual way of accessing the website, it is not effective for retaining users. Second, the search engine is not frequently used in isolation. That means that users visit another sections after checking the results of a search. The reason for this could be that search results are not useful to users and they need to visit another section to find the products they are looking for or it could be that after finding the desired products they visit different sections to look for related products.

Proposed actions Mechanisms to retain users that only visit specific *collections* should be included. Additionally, the marketing, mailing and advertising campaigns that are originating accesses to these sections should be redesigned since this section is not being effective.

C. BEHAVIORAL PATTERNS RELATED TO THE BUYING PROCESS

Regarding the buying process there are two specific actions that we are interested in. First, user sessions showing interest in acquiring a specific product. That corresponds to the events of *adding a product to the cart* and *adding a product to the wishlist*. In this regard, it is important to identify the sections visited just before such events happen. This way we could identify those sections that help users to find interesting products allowing to correlate such information with different access patterns. Second, sessions that buy some products, that is, sessions where the event *Buy products in the cart* happens. In this regard, it is important to analyse the relation between showing interest in a product and purchasing it.

First, we are interested in knowing from which sections the products are added to the cart or to the wishlist.

Query 5 *Which are the sections visited just before adding a product to the wishlist or to the cart?*

The query "$\Diamond$(?OR_V_BAR $\wedge$?M1) $\wedge$ $\bigcirc$(($\neg$?OR_VW_BAR) $\cup$ (Add_product_to_the_cart $\vee$ Add_product_to_the_wishlist)))" is performed for main sections and analogous one replacing (?OR_V_BAR $\wedge$?M1) by (?OR_W_BAR $\wedge$?S1) for secondary ones. Thereby, the query identifies for each main section ($m \in M^1$) and each secondary section ($s \in S1$) whether products are added to the cart or to the wishlist from that section m or s or from some of their subsections.

TABLE 6. Sections that lead to adding products to the wishlist or to the cart.

	Section	Total %	Relative %	Interest rate (%)
Main	Decorate	2.27	14.58	20.68
	Albums	0.23	1.48	6.32
	Stamp	3.35	21.52	22.01
	Tools	4.31	27.67	18.10
	Home decor-diy	0.67	4.30	9.57
	Papers	3.63	23.34	20.74
	Project life-smash	0.51	3.28	10.28
	Gifts	0.04	0.23	1.91
Secondary	Collections	0.72	4.61	3.96
	Designers	0.22	1.43	18.71
	Brands	1.38	8.87	14.72
	New products	1.43	9.17	11.17
	Offers	0.61	3.95	10.92
	Search results	3.10	19.93	13.96
	Thematics	0.45	2.86	13.50

Results Results are shown in Table 6 summarizes the results of the query by showing for each event the percentage of sessions meeting the query and the relative percentage of sessions with respect to the total number of sessions where some product is added to the wishlist or to the cart. Furthermore, this table shows in its last column the ratio between the number of accesses to the section that leads to showing interest in a product and the total number of accesses to the section. This so-called *interest rate* allows us to identify if accessing to a specific part of the website improves the probability of showing interest in a product. Note that in a session several products can be added from different web sections.

Interpretation There are four main sections that are leading to the addition of products to the wishlist or to the cart: *tools*, *papers*, *stamp* and *decorate*, in such order. Next, the use of the search engine is also a common pattern that leads to show interest in products. This is exemplified by both the total and the relative percentages. Regarding the interest rate, the previous sections also show a high interest rate, around 20%. A remarkable finding is that the *designers* secondary section has one of the highest interest rate. This could be explained by users that show fidelity to specific designers and look for their products. However, this section shows a low relative percentage of total interest and it has few accesses. Another relevant fact is that offers and new products sections are not showing high rates. However, it would be expected to be one of the

most interesting sections for regular customers. Finally, an interesting finding is that *collections* section shows one of the lowest interest ratio. However, it is one of the sections that poses a remarkable number of isolated and exclusive accesses (see Table 4 and 5). That means that users are accessing the website to visit products from collections but they are not showing interest in purchasing them.

Proposed actions It would be interesting to promote the *designers* section in order to facilitate the access to the section with the goal of increasing the number of users visiting such section. Similarly, it would be desirable to recommend products from the same designer when a specific one is being visited. This actions could increase the average interest rate of the website. Additionally, *offers*, *new products* and *collections* are not attracting the interest that could be expected based on the purpose of the first two sections and the high number of exclusive and first choice access to the last section. Therefore, marketing policies regarding them should be modified.

In order to complete the analysis, let us check the correlation between the previous results and actual purchases.

Query 6 For that, we have repeated the previous query adding that in the future the event *Buy products in the cart* happens, that is, adding "$\wedge \bigcirc \Diamond$ Buy_products_in_the_cart" to the query. This way sessions where the previous patterns happen and some products are later purchased are identified.

TABLE 7. Sections from where products are added to the cart or the wishlist in sessions that make a purchase.

	Section	Total %	Relative %	Purchase-interest rate (%)
Main	Decorate	0.73	18.12	32.19
	Albums	0.07	1.85	32.33
	Stamp	0.97	24.02	28.92
	Tools	1.32	32.85	30.75
	Home decor-diy	0.18	4.45	26.79
	Papers	1.01	25.14	27.90
	Project life-smash	0.16	3.86	30.56
	Gifts	0.02	0.41	46.15
Secondary	Collections	0.16	3.97	22.31
	Designers	0.06	1.54	27.81
	Brands	0.34	8.45	24.69
	New products	0.22	5.40	15.26
	Offers	0.13	3.33	21.83
	Search results	0.96	23.91	31.08
	Thematics	0.12	2.88	26.05

Results Table 7 shows the web sections from where products are added to the cart or the wishlist in sessions that make a purchase. Second column of the table show the total percentage of sessions showing this behavior. Third column of the table show the relative percentage with respect to the total number of sessions including a purchase action. Fourth column show the so-called *purchase-interest rate* that indicates the percentage of sessions that show interest in products and purchase them compared to the total number of sessions that show interest in products.

Interpretation Regarding the total and relative percentages of purchases, the results are in line with the ones shown in Table 6. The purchase-interest rate shows more interesting findings. There is not a preferred pattern at the global level to purchase a product since both main and secondary sections show a purchase-interest rate around 25% for both levels. Regarding main sections, the *gifts* section shows the highest rate. This could be due to the particularity of this section: it only provides initiation and gift packs. The remaining sections show a similar percentage and there is not a significant difference between sections attracting more interest and those with less. This is noticeable since it indicates that when users are interested in products there are not relevant patterns considering the section to which the product belongs. Regarding secondary sections, the *search engine* has the highest purchase-interest rate. This indicates that when people finds an interesting product through the search engine it is more likely to buy it. It is also remarkable that sections showing new products present a rate much lower than other web sections. This issue remarks the lack of effectiveness of this section and the need of improving it to increase its importance within the website.

Proposed actions Regarding main sections, it could be interesting to promote those sections that currently have less interest for users since they show similar purchasing rates when compared to the most visited sections. Some marketing strategy, such as discounts, should be considered to increase the purchase-interest rate of *new products* section since right now users add new products to the cart or to the wishlist but rarely buy them during the session. Finally, the use of the *search engine* should be promoted since it improves the probability of users buying products.

D. EVALUATING THE ANALYSIS EXECUTION TIME

Let us briefly analyse the execution time of the previous queries using the developed model checker. To obtain these measurements an Intel Core i7-4790K CPU at 4.00 GHz server has been used. The server runs an Ubuntu 16.04.1 LTS operating system. The model checker was executed using a single core.

TABLE 8. Execution time and number of queries executed for the analysis carried out.

Query type	Number of queries	Execution time (s)
Query 1	23	56.48
Query 2	51	128.84
Query 3	15	39.97
Query 4	15	43.51
Query 5	15	39.13
Query 6	15	39.72
Total	134	347.65

Table 8 summarizes the execution time of the six types of queries proposed to analyse the Up&Scrap web logs.

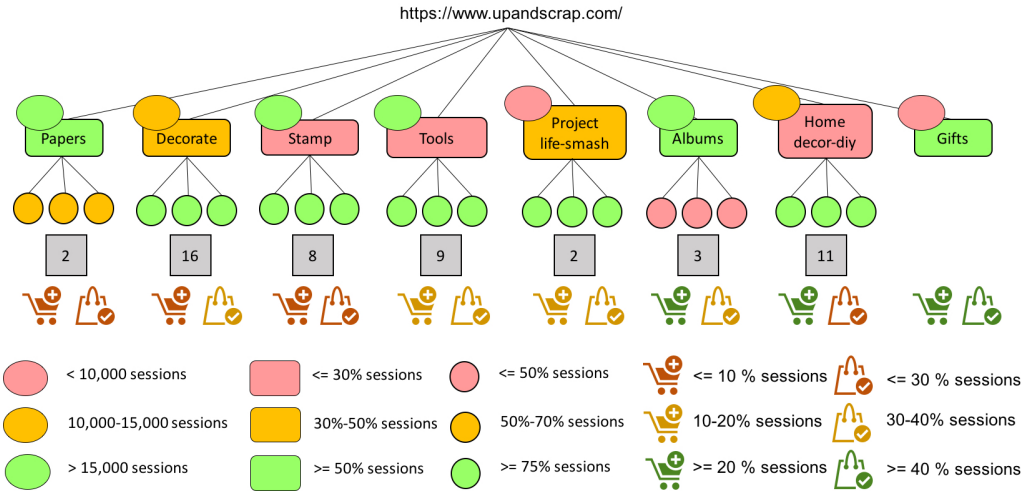

FIGURE 7. Summary of accesses to the main categories of the Up&Scrap website and their importance within the buying process.

As explained in Section III-C, the appearance of a variable in a query implies that a set of queries must be evaluated by considering all the possible values of the variable. Therefore, the second column of Table 8 shows the number of concrete queries executed for each type of query and the third column indicates the total execution time required.

As it has been shown, the number of concrete queries executed for each type of query depends on the website structure. Queries 3, 4, 5 and 6 evaluate 15 concrete queries because they analyse the behavior of the 8 level-1 main sections and the 7 level-1 secondary sections, for instance. In general, concrete queries present a similar performance with an average execution time of approximately 2.5 seconds per query. In total, the analysis of the Up&Scrap web logs includes 134 queries and it requires less than 6 minutes. Therefore, the amount of time required for the analysis performed is very reasonable according to the volume of data considered and the log complexity.

The execution time could be reduced by executing the model checker in parallel. In this regard, two complimentary approaches could be adopted: the log could be split into several parts which can be queried in parallel, or different queries could be executed in parallel since they are totally independent.

E. RESULTS VISUALIZATION AND INTERPRETATION

Previous section has shown the results of many interesting queries about the behavior of Up&Scrap customers. These results are presented using a tabular format. Nevertheless, this format hinders their understanding and requires of a deep analysis. Furthermore, analysing together the results from different queries could be interesting to extract meaningful conclusions and to identify different behavioral patterns. Therefore, a graphical representation can be used to present raw results in a more simple, intuitive and useful manner. With such purpose, in this section we propose to visualize query results according to the website structure and the product categorization. Thereby, different colours are used to present the query results depending if data is considered low, moderate or high. In this case, this distinction has been performed by analysing data in conjunction with domain experts. Let us going to show two example of visualization used to represent the results of the previous queries.

Figure 7 illustrates the most relevant access patterns to main sections and their importance within the buying process. The figure shows data presented using the product main categorization for the specific case of the Up&Scrap website (it corresponds to the left part of Figure 2). Data in the figure correspond to queries 1, 5 and 6 shown in the previous section. Thus, the figure shows the name of each level-1 section and the number of level-2 sections belonging to each level-1 section. Ovals in the figure show the total number of accesses to a section or some of its subsections. The colour in rectangles and circles indicates the percentage of sessions that access to such section. Finally, the colour of the shopping cart and the shopping bag indicates the interest rate and the purchase-interest rate, respectively. Note the figure does not include the specific value for the data considered for readability reasons, but this information could be easily incorporated.

The figure allow the user to obtain some interesting findings easily. For instance, the *tools* section is one of the most visited sections (green oval). Their visitors prefer level-2 categories over level-1 ones (red rectangle indicates a low access to level-1 categories and green circles shows a high number of accesses to the 9 level-2 categories). That means that when accessing to the tools section users prefer to look for specific products. However, the tools section only has moderate interest and purchase-interest rates (orange shopping cart and bag). So, despite of the large number of accesses

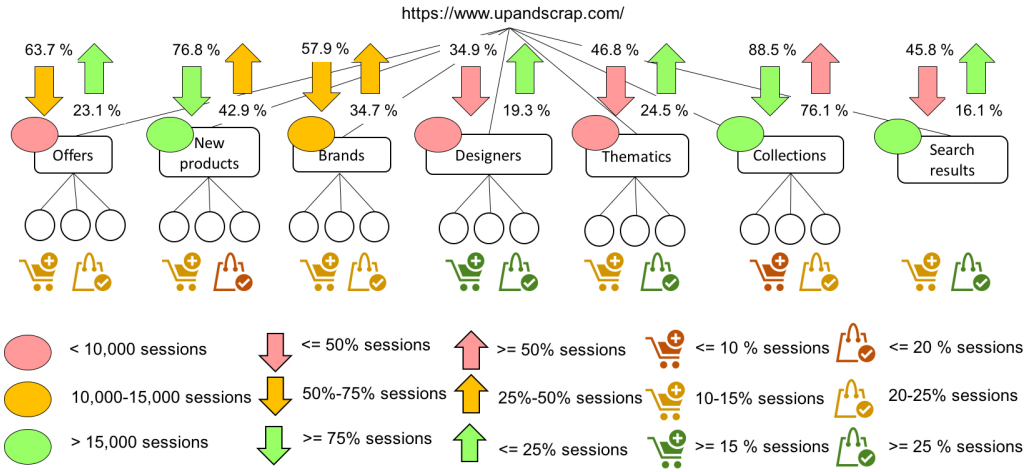

FIGURE 8. Summary of exclusive and first choice accesses to secondary sections of the Up&Scrap website and their importance within the buying process.

users are not more willing to buy products from this section. On the contrary, the *gifts* section is barely visited but it has high interest and purchase-interest rates. Finally, the *project life-smash* section is the unique one with a low number of subsections where users access mainly to subsections.

As a second example, Figure 8 analyses the exclusive and first choice accesses to secondary sections and their importance within the buying process. The figure presents data over the diagram that shows the secondary categorization of the Up&Scrap website (it corresponds to the right part of Figure 2). Data in the figure correspond to queries 1, 3, 4, 5 and 6. As in the previous example, ovals in the figure show the total number of accesses to a secondary section or some of its subsections. Arrows pointing down indicate the percentage of times that the session is visited as the user first choice. Arrows pointing up indicate the percentage of times that the session is visited in isolation, that is, the number of times that users only visit that section when navigating through the website. Finally, the colour of the shopping cart and the shopping bag indicates the interest rate and the purchase-interest rate, respectively.

Some interesting issues can be detected by analysing the previous figure. The *collections* section has a high number of sessions where this section is visited first by users (green downwards arrow) but it has a low number of users remaining in the website afterwards (the number of sessions visiting this section in isolation is high as shown by the red upwards arrow). Furthermore, products are not being added from the cart in these sessions (red shopping cart) but, when added, the probability of buying them is in the website average (orange shopping bag). The opposite behavior is shown in *thematics* and *designers* sections and the *search engine*. Finally, we can identify the *designers* sections as the most effective section with regard to the buying process, the *search engine* and the *thematics* section as two quite effective user choices and the *new products* section as the least effective one.

Finally, we would like to remark that the analysis presented in this paper has focused on detecting navigational and usage patterns and users' interest and preferences paying special attention to the buying process. One of the keys that has enabled this analysis is the flexibility and expressiveness of the model checking approach used and the tool implemented. This has allowed us to perform complex queries in order to identify non trivial behavioral patterns. In any case, the same approach could be used to analyse other type of patterns by following the same methodological approach shown during this section.

VII. CONCLUSIONS AND FUTURE WORK

In the case of open systems, where the sequences of interactions (stored as system logs) are not constrained by a workflow, process mining techniques whose objective is to extract a process model will usually provide with either overfitting *spaghetti* models or underfitting *flower* models, from which little interesting information can be extracted.

A more flexible approach is required. In the paper we apply LTL-based model checking techniques to analyse e-commerce web logs. To enable this analysis, we have proposed a common way of representing event types and attributes considering the e-commerce web structure, the product categorization and the possibilities of users to navigate throw the website according to such organization.

From this structural point of view, the paper proposes a set of query patterns, translated into LTL formulas, which are of interest for the domain of electronic commerce The answers to the queries, in terms of the number (or percentage) of traces satisfying the corresponding formula, allows to extract interesting correlations among sequences of events, which can be interpreted in terms of users' behavior. Among the wide set of possible behaviors, we have concentrated on finding how the different website sections are visited and which navigational patterns are related to buying actions.

As a use case the paper presents the application of the approach to the Up&Scrap e-commerce website. The analysis carried out has allowed us to identify several issues and to propose improvements regarding the product categorization and the organization of some of the website sections, which have been transferred to the enterprise managers.

Although the paper is strongly related to that website, the proposed approach is general and the methodology is applicable to structured e-commerce websites. The first phase of the methodology, the preprocessing phase, is the one which is specific for each e-commerce website, since it depends on the specific system log and, meanwhile the analysis technique and the queries can be completely reused.

On the other hand, the analysis in the paper has been made for a log corresponding to two months of use. However, the proposed method is directly applicable to much bigger logs, since both the method and the tool scale very well: it can be executed in parallel, deploying different parallel servers with different parts of the log and executing the queries in parallel.

As a near future work we want to provide the analysis tool with a graphical interface, for both the input of properties to be analysed and the output of results, with the aim of facilitating its use for non-technical staff, providing with an abstraction level hiding the LTL formalism. We also plan to extend the set of studied patterns in order to analyse more behavioral patterns and to facilitate their automatic discovery. For that, a side-by-side work with specialists of the problem domain is required in order to define a set of interesting queries as wide as possible. Additionally, extending the web server logs with information about users or online customer reviews is going to be studied. User's information would allow us to study multi session patterns and correlate results with demographic information; while, online reviews would allow us to analize customer's feedbacks in order to recommend products (some statistical models are presented in [52], [53]). Finally, we would like to consider the extension of the approach to consider time constraints between the events in the line shown in [52] and [53].

ACKNOWLEDGEMENTS

The authors of this paper want to specially thank the Up&Scrap team for their collaboration, for providing the data used in this study, and for giving feedback on the results.

REFERENCES

[1] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications," *Data Mining Knowl. Discovery*, vol. 5, no. 1, pp. 115–153, Jan. 2001.

[2] N. Poggi, D. Carrera, R. Gavalda, J. Torres, and E. Ayguadé, "Characterization of workload and resource consumption for an online travel and booking site," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Apr. 2010, pp. 1–10.

[3] R. Kohavi, "Mining e-commerce data: The good, the bad, and the ugly," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 8–13.

[4] W. W. Moe and P. S. Fader, "Dynamic conversion behavior at e-commerce sites," *Manage. Sci.*, vol. 50, no. 3, pp. 326–335, 2004.

[5] G. Liu *et al.*, "Repeat buyer prediction for e-commerce," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2016, pp. 155–164.

[6] J. D. Xu, "Retaining customers by utilizing technology-facilitated chat: Mitigating website anxiety and task complexity," *Inf. Manage.*, vol. 53, no. 5, pp. 554–569, 2016.

[7] 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, 2011.

[8] R. Kosala and H. Blockeel, "Web mining research: A survey," *SIGKDD Explor. Newsl.*, vol. 2, no. 1, pp. 1–15, Jun. 2000.

[9] F. M. Facca and P. L. Lanzi, "Mining interesting knowledge from Weblogs: A survey," *Data Knowl. Eng.*, vol. 53, no. 3, pp. 225–241, 2005.

[10] C. J. Carmona and S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. García, "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11243–11249, 2012.

[11] Q. Song and M. Shepperd, "Mining Web browsing patterns for e-commerce," *Comput. Ind.*, vol. 57, no. 7, pp. 622–630, 2006.

[12] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. M. Pérez, and I. Perona, "Web usage and content mining to extract knowledge for modelling the users of the bidasoa turismo website and to adapt it," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7478–7491, 2013.

[13] J. K. Gerrikagoitia, I. Castander, and F. Rebön, and A. Alzua-Sorzabal, "New trends of intelligent e-marketing based on Web mining for e-shops," *Procedia-Social Behavioral Sci.*, vol. 175, pp. 75–83, Sep. 2015.

[14] Y. H. Cho and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Syst. Appl.*, vol. 26, no. 2, pp. 233–246, 2004.

[15] K.-J. Kim and H. Ahn, "A recommender system using GA K-means clustering in an online shopping market," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 1200–1209, 2008.

[16] 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, 2015.

[17] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from Web data," *SIGKDD Explor. Newsl.*, vol. 1, no. 2, pp. 12–23, Jan. 2000.

[18] Q. Zhang and R. S. Segall, "Web mining: A survey of current research, techniques, and software," *Int. J. Inf. Technol. Decision Making*, vol. 7, no. 4, pp. 683–720, 2008.

[19] B. Singh and H. K. Singh, "Web data mining research: A survey," in *Proc. IEEE Int. Conf. Comput. Int. Comput. Res. (ICCIC)*, Sep. 2010, pp. 1–10.

[20] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 1st ed. Heidelberg, Germany: Springer-Verlag, 2011.

[21] N. Poggi, V. Muthusamy, D. Carrera, and R. Khalaf, "Business process mining from e-commerce Web logs," in *Proc. 11th Int. Conf. Bus. Process Manage.*, Berlin, Germany, 2013, pp. 65–80.

[22] F. M. Maggi, R. P. J. C. Bose, and W. M. P. van der Aalst, *Efficient Discovery of Understandable Declarative Process Models from Event Logs*. Berlin, Germany, 2012, pp. 270–285.

[23] M. Räim, C. Di Ciccio, F. M. Maggi, M. Mecella, and J. Mendling, "Log-based understanding of business processes through temporal logic query checking," in *Proc. Move Meaningful Internet Syst. (OTM), Conf. Confederated Int. CoopIS*, Amantea, Italy, Oct. 2014, pp. 75–92.

[24] A. Burattin, M. Cimitile, F. M. Maggi, and A. Sperduti, "Online discovery of declarative process models from event streams," *IEEE Trans. Services Comput.*, vol. 8, no. 6, pp. 833–846, Jun. 2015.

[25] A. Duret-Lutz, A. Lewkowicz, A. Fauchille, T. Michaud, E. Renault, and L. Xu, "Spot 2.0—A framework for LTL and ω-automata manipulation," in *Proc. 14th Int. Symp. Automated Technol. Verification Anal. (ATVA)*, Oct. 2016, pp. 122–129.

[26] 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.

[27] L. G. Vasconcelos, R. D. C. Santos, and L. A. Baldochi, "Exploiting client logs to support the construction of adaptive e-commerce applications," in *Proc. IEEE 13th Int. Conf. e-Bus. Eng. (ICEBE)*, Apr. 2016, pp. 164–169.

[28] Y.-L. Chen, M.-H. Kuo, S.-Y. Wu, and K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data," *Electron. Commerce Res. Appl.*, vol. 8, no. 5, pp. 241–251, Oct. 2009.

[29] S. Kim, J. Yeo, E. Koh, and N. Lipka, "Purchase influence mining: Identifying top-k items attracting purchase of target item," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 57–58

[30] S. D. Bernhard, C. K. Leung, V. J. Reimer, and J. Westlake, "Clickstream prediction using sequential stream mining techniques with Markov Chains," in *Proc. 20th Int. Database Eng. Appl. Symp.*, New York, NY, USA, 2016, pp. 24–33.
[31] L. Lu, M. Dunham, and Y. Meng, "Mining significant usage patterns from clickstream data," in *Proc. 7th Int. Conf. Knowl. Discovery Web Adv. Web Mining Web Usage Anal.*, Berlin, Germany, 2006, pp. 1–17.
[32] W. M. van Der Aalst, M. Pesic, and H. Schonenberg, "Declarative workflows: Balancing between flexibility and support," *Comput. Sci.-Res. Develop.*, vol. 23, no. 2, pp. 99–113, 2009.
[33] A. Burattin, F. M. Maggi, and A. Sperduti, "Conformance checking based on multi-perspective declarative process models," *Expert Syst. Appl.*, vol. 65, pp. 194–211, Oct. 2016.
[34] (2017). *Google Analytics*. accessed May 22, 2017. [Online]. Available: https://analytics.google.com/analytics/web/
[35] (2017). *Clicky*. accessed May 22, 2017. [Online]. Available: https://clicky.com
[36] (2017). *Piwik Open-Source Analytics Platform*. accessed May 22, 2017. [Online]. Available: https://piwik.org
[37] (2017). *Adobe Analytics*. accessed May 22, 2017. [Online]. Available: https://analytics.google.com/analytics/web/
[38] (2017). *W3counter*. accessed May 22, 2017. [Online]. Available: https://www.w3counter.com
[39] A. Pnueli and Z. Manna, *The Temporal Logic of Reactive and Concurrent Systems*. New York, NY, USA: Springer-Verlag, 1992.
[40] E. M. Clarke, A. Emerson, and A. P. Sistla, "Automatic verification of finite state concurrent system using temporal logic specifications: A practical approach," in *Proc. 10th ACM SIGACT-SIGPLAN Symp. Principles Program. Lang. (POPL)*, Austin, TX, USA, Jan. 1983, pp. 117–126.
[41] E. Clarke, O. Grumberg, and D. Long, "Verification tools for finite-state concurrent systems," in *Proc. Workshop/School/Symp. REX Project (Res. Edu. Concurrent Syst.)*, 1993, pp. 124–175.
[42] J. Couvreur, "On-the-fly verification of linear temporal logic," in *Proc. Formal Methods, World Congr. Formal Methods Develop. Comput. Syst.*, Toulouse, France, Sep. 1999, pp. 253–271.
[43] W. van der Aalst and M. Pesic, "Decserflow: Towards a truly declarative service flow language," in *Proc. Role Bus. Process. Service Oriented Archit.*, Dagstuhl, Germany, 2006, pp. 1–23.
[44] A. Bauer and P. Haslum, "LTL goal specifications revisited," in *Proc. 19th Eur. Conf. Artif. Intell. Conf. ECAI*, Amsterdam, The Netherlands, 2010, pp. 881–886.
[45] G. De Giacomo, R. De Masellis, and M. Montali, "Reasoning on LTL on finite traces: Insensitivity to infiniteness," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1027–1033.
[46] P. Álvarez, J. Fabra, S. Hernández, and J. Ezpeleta, "Alignment of teacher's plan and students' use of LMS resources. Analysis of Moodle logs," in *Proc. 15th Int. Conf. Inf. Technol. Based Higher Edu. Training (ITHET)*, Sep. 2016, pp. 1–8.
[47] Common Log Format (CLF). (1995). *The World Wide Web Consortium (W3C)*. [Online]. Available: http://www.w3.org/Daemon/User/Config/Logging.html#common-logfileforma
[48] M. Srivastava, R. Garg, and P. Mishra, "Preprocessing techniques in Web usage mining: A survey," *Int. J. Comput. Appl.*, vol. 97, no. 18, p. 1, 2014.
[49] K. S. Reddy, M. K. Reddy, and V. Sitaramulu, "An effective data preprocessing method for Web usage mining," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, 2013, pp. 7–10.
[50] G. Neelima and S. Rodda, "Predicting user behavior through sessions using the Web log mining," in *Proc. Int. Conf. Adv. Human Mach. Interaction (HMI)*, 2016, pp. 1–5.
[51] R. Y. Lau, J. L. Zhao, G. Chen, and X. Guo, "Big data commerce," *Inf. Manage.*, vol. 53, no. 8, pp. 929–933, 2016.
[52] J. Qi, Z. Zhang, S. Jeon, and Y. Zhou, "Mining customer requirements from online reviews: A product improvement perspective," *Inf. Manage.*, vol. 53, no. 8, pp. 951–963, 2016.
[53] Y. Kang and L. Zhou, "Rube: Rule-based methods for extracting product features from online consumer reviews," *Inf. Manage.*, vol. 54, no. 2, pp. 166–176, 2017.

SERGIO HERNÁNDEZ received the Ph.D. degree in computer science engineering from the University of Zaragoza, Spain, in 2016. He is currently a Postdoctoral Researcher with the Department of Computer Science and Systems Engineering, University of Zaragoza. His main research areas focus on cluster, grid and cloud computing infrastructures, integration and interoperability of heterogeneous computing infrastructures, scientific workflows, and process and data mining techniques for the analysis of big data logs.

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 the University of Zaragoza since 2000. His research interests focus on the integration problems of network-based systems (cloud-based and service-based systems, mainly) and the use of novel techniques and methodologies for solving them, and the application of formal analysis techniques to mine event logs and databases (in the domain of e-commerce, e-learning, cybersecurity, or health, for example).

JAVIER FABRA received the Ph.D. in computer science from the University of Zaragoza, Spain, in 2010. He holds an associate professor position with the Department of Computer Science and Systems Engineering, University of Zaragoza, since 2008. His main research areas focus on service-oriented computing and architectures, semantic and scientific computing, and interoperability issues in cluster, grid, and cloud scenarios by means of the application of high-level Petri nets.

JOAQUÍN EZPELETA received the M.S. degree in mathematics and the Ph.D. degree in computer science from the University of Zaragoza, Spain. He was a Researcher with the Laboratory of Methods and Architectures for Information Systems, University of Paris-6, and the Digital Enterprise Research Institute, National University of Ireland, Galway. He is currently a Professor of the Department of Computer Science and Systems Engineering, University of Zaragoza, where he is involved in formal methods for sequential and concurrent programming and service-oriented architectures. He has co-authored more than 80 research papers and participated in numerous program comities of international conferences, being also a Reviewer of prestigious research journals. His research has focused on the problem of modeling, analysis, and control synthesis for concurrent systems, and the application of formal techniques to help in the development of correct distributed systems based on Internet and cloud technologies, and the parallel processing of data and computing intensive computing problems.