Using model checking to identify customers' purchasing behaviour in an e-commerce

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, Spain
{shernandez,alvaper,jfabra,ezpeleta}@unizar.es

Abstract. Understanding customers' behaviour is essential to adapt e-commerce websites to their preferences and requirements. Process mining techniques can be used to analyse e-commerce web logs with such purpose. However, procedural techniques are not suitable to analyse logs from open systems and current declarative ones focus on extracting the most frequent behavioural constraints. In this paper the use of a declarative approach based on Linear Temporal Logic and model checking is explored for the analysis of a real e-commerce website. This approach explores complex queries adapted to the problem domain. Results show that analysing both common and uncommon behavioural patterns provides valuable insights that can be used to adapt the website and to improve marketing and advertising campaigns.

Keywords: Linear temporal logic, model checking, declarative process mining, complex behavioural patterns, purchasing process

1 Introduction

The appearance of e-commerce websites has revolutionized the shopping context allowing people to purchase products from anywhere at any time. Understanding users' requirements, necessities and preferences is key to survive the competition and succeed [1]. Nevertheless, it is a complex task [2].

Users' behaviour when using an e-commerce website is recorded in the web server logs which store the sequence of requests performed by users. Analysing this type of logs to extract and identify behaviour has been the goal of both data and process mining communities. On the one hand, data mining approaches are focused on extracting behaviour by applying clustering, sequence pattern mining and rule association techniques [3,4,5,6]. However, these techniques do not consider the order of activities recorded in the log and, as a consequence, they lose information about which events influence the occurrence of other events, for instance. On the other hand, the process mining community mainly focus on discovering a procedural model from the log [7]. However, in the e-commerce domain this approach is not very suitable [1]. An e-commerce website is an open system where any behaviour is possible and a meaningful process model cannot

be identified since the application of procedural techniques usually provides a *flower* model or a *spaghetti* one from where no useful information can be extracted [7]. In this context, *declarative* approaches are more suitable since they allow to look for specific behavioural constraints [8,9].

In this paper we propose the use of Linear Temporal Logic (LTL) [10] and model checking [11] techniques to explore complex behavioural patterns in open systems as the case of the e-commerce domain. The approach is then illustrated by its application to the analysis of the Up&Scrap e-commerce website[1]. The goal is to analyse multi-perspective behavioural patterns related with the buying process defining and checking LTL formulas that describe such behaviour against the log model. By applying this kind of analysis, we are interested in finding causal dependencies among events and the fact of purchasing goods, that is, behavioural patterns whose occurrence increase the probability of purchase. The obtained information can be used to improve the website design and contents, to adapt and personalize contents or to recommend products with the ultimate goal of increasing sales.

The use of temporal logics to analyse event logs using a declarative approach has been previously explored [12,8,13,14,15,16,9,17]. Nevertheless, these approaches struggle with multi-perspective analysis that is and/or the use of ad-hoc complex queries. Most approaches do not consider the possibility or including data attributes to enable multi-perspective analysis [8,13,14,15,16], or they only allow specific perspectives [12]. This limits the effectiveness in the approach when the log contains many data attributes and their importance is key for the analysis as in the e-commerce domain where identifying and differentiating web sections is fundamental. Furthermore, many techniques are based on the use of the Declare language [8] and, as a consequence, they can only explore a reduced set of predefined patterns [8,13,14,15,9,17] which can be hardly extended [9,17]. This issue reduces the type of behavioural patterns and constraints that can be discovered or analysed causing that interesting behaviours remain unexplored although they could provide valuable information.

Compared with the above mentioned approaches, our proposal is more general since it enables multi-perspective analysis without limitations on the perspectives that can be used, it allows the use of any kind of LTL formula and it is able to handle large event logs as the one explored in this paper. Finally, it must be mentioned that the approach proposed in this paper focuses on supervised and semi-supervised mining while previous approaches have mainly focused in unsupervised one. Although supervised techniques require a business expert to define the behavioural patterns that must be explored, they have the advantage that uncommon but valuable patterns, as the ones related with the purchasing process, can be identified. In this regard, unsupervised techniques are less suitable since they focus on extracting the most frequent patterns.

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

2 Model checking-based analysis of the Up&Scrap logs

In this work we have applied the LTL-based model checking approach to the web logs of the Up&Scrap website. Up&Scrap is the leader company in Spain for the sale of equipment for scrapbooking. The company website is structured in a set of main sections which organize the products in eight categories (paper, decorate, stamp, tools, project life-smash, albums, home decor-diy and gifts) according to the product categorization that includes two levels of depth. There are also secondary sections which provide an alternative way of accessing products by brands, collections, thematics, designers, offers and new products. Additionally, a search engine is provided to look for products within the website.

We have analysed the web logs of two months which include $8,607,625$ HTTP requests. Initially, this information must be preprocessed to remove automatic, erroneous and irrelevant requests and to group requests belonging to the same session (*sessionization*). Next, each request must be analysed to identify interesting events for the analysis that is going to be carried out. Since our aim is to analyse the customers' purchasing behaviour, we have identified the following 12 types of events: *Visit_homepage, Visit_product, Visit_main_section_L1, Visit_main_section_L2, Visit_secondary_section_L1, Visit_secondary_section_L2, Buy_products, Delete_product_from_the_cart, Add_wishlist_products_to_the_cart, Add_product_to_the_wishlist, Add_product_to_the_cart, Update_product_from_cart.* Also, the category and subcategory being accessed are obtained and included in the final log along with the URL accessed, the operation code (GET or POST), the status code and the timestamp. After the preprocessing, the final log used for the analysis contains $1,331,697$ events corresponding to $144,330$ user sessions.

To enable the LTL-based model checking analysis, the set of traces (user sessions) in the log is considered as the model representing the process and each event is represented as the conjunction of atomic propositions corresponding to the event activity and its attributes. Figure 1 shows a simplified example of trace in the previous format showing the event activity and part of the attributes defining the different perspectives that can be used in the analysis. Therefore, LTL formulas can be used to analyse the trace behaviour using the model checker tool presented in [18] that is based on the use of the Spot libraries [19]. To exemplify the kind of behavioural patterns that can be discovered by using this technique, we are going to focus on the buying process.

3 Analysis of Up&Scrap customers' purchasing behaviour

To enable the analysis, we have defined a set of variables and macros in the model checker tool described in [18]. Their use improves the queries readability and allows the execution of a query on multiple values such as all the sections of the web page, for instance. The defined variables and macros are:

- Variable *?visit_main* includes the events referred to visit a main section. Values: Visit_main_section_L1 and Visit_main_section_L2.

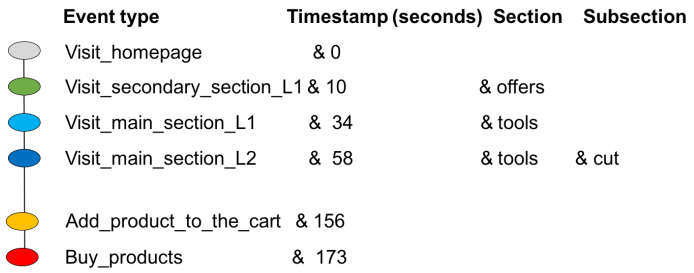

Fig. 1. Simplified example of a trace from the Up&Scrap log showing the activity and some data attributes as a logical conjunction.

- Variable *?visit_secondary* includes the events referred to visit a secondary section. Values: Visit_secondary_section_L1 and Visit_secondary_section_L2.
- Variable *?visit* includes events in *?visit_main* and *?visit_secondary*.
- Variable *?main_cat* includes all the possible values of main categories. Values: papers, decorate, stamp, tools, project life-smash, home decor-diy and gifts.
- Variable *?sec_cat* includes all the possible values of secondary categories. Values: offers, new products, collections, thematics, designers, brands and search results.
- Macros *?OR_VISIT_MAIN*, *?OR_VISIT_SEC*, *?OR_VISIT* correspond to a logical OR of "?visit_main", "?visit_secondary" and "?visit", respectively.

In the buying process there are two essential actions: adding products to the cart and buying the products in the cart. Our goal is to analyse specific sections that lead to the appearance of these two actions. In the Up&Scrap website, products can be reached both from main and secondary sections. Furthermore, products can be added to the cart from the own product page, which has no information about the section from where the product has been accessed, and from the product listings in the sections. Therefore, identifying web sections that lead to purchases is not immediate since the query must consider both possibilities. With such purpose we have executed two types of queries[2]:

1. *Which are the main sections from where products are added to the cart?*
   $\Diamond$(?OR_VISIT_MAIN $\wedge$?main_cat) $\wedge$ $\bigcirc((\neg$?OR_VISIT) $\cup$ (Add_product_to_the_cart)))
2. *Which are the main sections from where products are added to the cart in sessions that buy some product?*
   $\Diamond$(?OR_VISIT_MAIN $\wedge$?main_cat) $\wedge$ $\bigcirc((\neg$?OR_VISIT) $\cup$ (Add_product_to_the_cart)) $\wedge$ $\bigcirc\Diamond$ Buy_products)
3. *Which are the secondary sections from where products are added to the cart?*
   $\Diamond$(?OR_VISIT_SEC $\wedge$?sec_cat) $\wedge$ $\bigcirc((\neg$?OR_VISIT) $\cup$ (Add_product_to_the_cart)))

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

4. *Which are the secondary sections from where products are added to the cart in sessions that buy some product?*
$\Diamond(?\text{OR\_VISIT\_SEC} \wedge ?\text{sec\_cat}) \wedge \bigcirc((\neg ?\text{OR\_VISIT}) \cup (\text{Add\_product\_to\_the\_cart})) \wedge \bigcirc\Diamond \text{Buy\_products})$

Note the use of variables and macros allows to write more compact formulas. Nevertheless, macros must be substituted with their appropriate values and variables must be evaluated for each possible value. For example, one of the specific formulas evaluated in the last case for the *offers* secondary sections is: $\Diamond((\text{Visit\_secondary\_section\_L1} \vee \text{Visit\_secondary\_section\_L2}) \wedge \text{offers}) \wedge \bigcirc((\neg(\text{Visit\_main\_section\_L1} \vee \text{Visit\_main\_section\_L2} \vee \text{Visit\_secondary\_section\_L1} \vee \text{Visit\_secondary\_section\_L2})) \cup (\text{Add\_product\_to\_the\_cart})) \wedge \bigcirc\Diamond \text{Buy\_products})$

Table 1 summarizes the results. For the first query type (queries 1 and 3) we define the so-called *interest-rate* as the ratio between the number of accesses to the section that leads to adding a product to the cart and the total number of accesses to the section. For the second query type (queries 2 and 4) we introduce the so-called *purchase-interest rate* as the percentage of sessions that add products to the cart and purchase them compared to the total number of sessions that add products to the cart. These metrics are inspired in known concepts, as support or confidence, commonly used in other declarative approaches [8,9].

Table 1. Web sections that lead to adding products to the cart and buying products. Green, white and red cells show high, medium or low rate respectively. Additionally, upwards and downwards arrows also show a high and low rate, respectively.

	Section	Interest rate (%)	Purchase-interest rate (%)
Main	Decorate	20.68 ↑	32.19
	Albums	6.32 ↓	32.33
	Stamp	22.01 ↑	28.92
	Tools	18.10 ↑	30.75
	Home decor-diy	9.57	26.79
	Papers	20.74 ↑	27.90
	Project life-smash	10.28	30.56
	Gifts	1.91 ↓	46.15 ↑
Secondary	Collections	3.96 ↓	22.31 ↓
	Designers	18.71 ↑	27.81
	Brands	14.72	24.69
	New products	11.17	15.26 ↓
	Offers	10.92	21.83 ↓
	Search results	13.96	31.08
	Thematics	13.50	26.05

There are four main sections that are leading to the addition of products to the cart according to their interest rate: *stamp*, *papers*, *decorate* and *tools*. Regarding secondary sections, a remarkable finding is that the *designers* section has one of the highest interest rate. This is explained by users that show fidelity to specific designers and look for their products. Another relevant discovery is that *offers* and *new products* sections are not showing high interest rates. Regarding the purchase-interest rate, surprisingly,

the *gifts* section has the lowest interest rate but the highest purchase-interest one. The remaining main sections show a similar percentage and there are not a significant differences between sections. On the contrary, the *search engine* has the highest purchase-interest rate among secondary ones. 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 *new products* and *offers* sections present a rate much lower than other web sections. This issue points towards the lack of effectiveness of these sections and the need of improving its importance within the website.

4 Conclusions and future work

The analysis of event logs of open systems, as the ones recorded in the web server logs of e-commerce websites, requires the use of flexible declarative approaches able to identify valuable behavioural patterns. Current approaches present limitations in the use of data attributes and complex user-defined patterns. In this paper we have proposed a LTL-based model checking approach that is able of effectively analyse this type of logs by exploring complex domain-related behavioural patterns that include data attributes and allow multi-perspective analysis. Analysis results provide insights on most and less effective sections regarding the buying process that have been used to improve the Up&Scrap website design.

As future work, we intend to develop a methodology for the analysis of any e-commerce website using this approach and to apply the technique in other application domains.

Acknowledgements

This work has been supported by the TIN2014-56633-C3-2-R research project, granted by the Spanish Ministerio de Economía y Competitividad and teh UZCUD-2016-TEC-06 research project granted by the University of Zaragoza. 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. N. Poggi, V. Muthusamy, D. Carrera, R. Khalaf, Business process mining from e-commerce web logs, in: Proceedings of the 11th International Conference on Business Process Management, BPM'13, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 65–80. doi:10.1007/978-3-642-40176-3_7.
2. R. Kohavi, Mining e-commerce data: the good, the bad, and the ugly, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001, pp. 8–13.
3. Q. Zhang, R. S. Segall, Web mining: a survey of current research, techniques, and software, International Journal of Information Technology & Decision Making 7 (04) (2008) 683–720. doi:10.1142/S0219622008003150.
4. Q. Su, L. Chen, A method for discovering clusters of e-commerce interest patterns using click-stream data, Electronic Commerce Research and Applications 14 (1) (2015) 1 – 13. doi:10.1016/j.elerap.2014.10.002.

5. S. Kim, J. Yeo, E. Koh, N. Lipka, Purchase influence mining: Identifying top-k items attracting purchase of target item, in: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 57–58.
6. J. D. Xu, Retaining customers by utilizing technology-facilitated chat: Mitigating website anxiety and task complexity, Information & Management 53 (5) (2016) 554 – 569. doi:10.1016/j.im.2015.12.007.
7. W. M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, 1st Edition, Springer Publishing Company, Incorporated, 2011. doi:10.1007/978-3-642-19345-3.
8. W. M. van Der Aalst, M. Pesic, H. Schonenberg, Declarative workflows: Balancing between flexibility and support, Computer Science-Research and Development 23 (2) (2009) 99–113. doi:10.1007/s00450-009-0057-9.
9. A. Burattin, F. M. Maggi, A. Sperduti, Conformance checking based on multi-perspective declarative process models, Expert Systems with Applications 65 (2016) 194 – 211. doi:10.1016/j.eswa.2016.08.040.
10. A. Pnueli, Z. Manna, The temporal logic of reactive and concurrent systems (1992). doi:10.1007/978-1-4612-0931-7.
11. E. Clarke, O. Grumberg, D. Long, Verification tools for finite-state concurrent systems, in: Workshop/School/Symposium of the REX Project (Research and Education in Concurrent Systems), Springer, 1993, pp. 124–175.
12. W. M. van der Aalst, H. De Beer, B. F. van Dongen, Process mining and verification of properties: An approach based on temporal logic, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer, 2005, pp. 130–147.
13. F. M. Maggi, R. P. J. C. Bose, W. M. P. van der Aalst, Efficient Discovery of Understandable Declarative Process Models from Event Logs, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 270–285. doi:10.1007/978-3-642-31095-9_18.
14. C. D. Ciccio, M. Mecella, A two-step fast algorithm for the automated discovery of declarative workflows, in: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2013, pp. 135–142. doi:10.1109/CIDM.2013.6597228.
15. M. Westergaard, C. Stahl, H. A. Reijers, UnconstrainedMiner: efficient discovery of generalized declarative process models, BPM Center Report BPM-13-28, BPMcenter. org (2013) 28.
16. M. Räim, C. Di Ciccio, F. M. Maggi, M. Mecella, J. Mendling, Log-based understanding of business processes through temporal logic query checking, in: On the Move to Meaningful Internet Systems: OTM 2014 Conferences: Confederated International Conferences: CoopIS, and ODBASE 2014, Proceedings, Springer Berlin Heidelberg, 2014, pp. 75–92. doi:10.1007/978-3-662-45563-0_5.
17. S. Schönig, C. Di Ciccio, F. M. Maggi, J. Mendling, Discovery of Multi-perspective Declarative Process Models, Springer International Publishing, Cham, 2016, pp. 87–103. doi:10.1007/978-3-319-46295-0_6.
18. P. Álvarez, J. Fabra, S. Hernández, J. Ezpeleta, Alignment of teacher's plan and students' use of lms resources. analysis of moodle logs, in: 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET), 2016, pp. 1–8. doi:10.1109/ITHET.2016.7760720.
19. A. Duret-Lutz, D. Poitrenaud, Spot: an extensible model checking library using transition-based generalized b uuml;chi automata, in: Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, 2004. (MASCOTS 2004). Proceedings. The IEEE Computer Society's 12th Annual International Symposium on, 2004, pp. 76–83. doi:10.1109/MASCOT.2004.1348184.