Tracking of User's behaviour in Structured E-Commerce Website

Chandana M.
UG Student
Department of Computer Science & Engineering
Sapthagiri College of Engineering, Bangalore, India

Komal Kumari
UG Student
Department of Computer Science & Engineering
Sapthagiri College of Engineering, Bangalore, India

Madduri Uma
UG Student
Department of Computer Science & Engineering
Sapthagiri College of Engineering, Bangalore, India

Smriti Gupta
UG Student
Department of Computer Science & Engineering
Sapthagiri College of Engineering, Bangalore, India

Gajendra Prasad K. C.
Assistant Professor
Department of Computer Science & Engineering
Sapthagiri College of Engineering, Bangalore, India

Abstract

Online shopping is becoming more and more common in our daily lives. Tracking user's interests and behaviour is essential in order to fulfil customer's requirements. The information about user's behaviour is stored in the web server logs. Absorbing a view of the process followed by user's during a session can be of great interest to identify the behavioural patterns. The analysis of such information has focused on applying data mining techniques. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. It is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. To address this issue, this paper proposes a linear temporal logic model checking method for the analysis of structured e-commerce web logs. By defining a common way of tracing log records according to the ecommerce structure, web logs can be converted into event logs where the behaviour of user's is tracked. Then, different predefined queries can be performed to identify different actions performed by a user during a session. 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.

Keywords: Data Mining, E-Commerce, Web Logs Analysis, behavioural Patterns, Model Checking

I. INTRODUCTION

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. It is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. In today's world, the way of shopping has changed. People are buying more and more over the Internet instead of going traditional shopping. E-commerce provides customers with the facilities of browsing endless product catalogues, comparing prices, creating wishlist and enjoying a better service based on their individual interests. 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. E-commerce business analysts require to know and understand consumer's behaviour when those go through the website, as well as trying to identify the reasons that motivated them to purchase. Getting this behavioural knowledge will allow e-commerce websites to deliver a more personalized service to customers and increasing benefits. Usually, the user's activities are recorded in the web server logs. Web server logs are stored in an ordered way, the sequence of web events generated by each user is stored as a separate log. The very valuable user's behaviour is hidden in these logs, which must be discovered and analysed. In 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. The focus is on the user's interest in the different product categories and their characterization consist of the list of visited categories and the frequency of such visits. The goal is to analyse the usage of e-commerce websites and to discover customer's complex behavioural patterns by means of checking temporal logic formulas describing such behaviours against the log model. The business analyst can use a set of temporal logic patterns to formulate queries that could help him to discover and understand the way clients use the website. Customer's reviews are considered for the further improvements.

Customer can give online reviews about the purchase and that can be considered for the further improvement of product and services. As a use case of the proposed approach we describe the analysis carried out for the Up & Scrap e-commerce website, an important on-line Spanish provider of scraping products. The case study describes the way raw logs have been processed, how the traces have been extracted, how user's behavioural patterns have been formulated and checked against the log.

II. RELATED WORK

In the field of e-commerce, most data mining techniques process server logs to extract the sequences of user navigation events. In previous approaches, 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, which tries to identify the user's interests from the contents of the visited pages. This information can subsequently be used to improve the website contents and structure. Another researcher applies alternative mining techniques to predict the user's behaviour. 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 [3]. These approaches having drawbacks: the process of creating these models is expensive, besides, the model does not have information to know how the current navigational state and how future states representing long-term goals can be reached. Alternative to data mining, process mining techniques try to obtain causal relations between the events of user's sessions. This technique used can only consider events with a very high abstraction level having difficulties in the identification of patterns as in the case of buying events, a set of constraints is assumed, and 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 is used, defined by the Declare property description language. MP-Declare extends Declare by introducing the possibility of defining data and time constraints in the Declare patterns. Another one is Google Analytics, controls the network traffic, collects information and displays reports about users' behaviour [4]. 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, such as. Dependence on. JavaScript and cookies, the necessity of adding page tags to every page, as a result, customers may experience a change in the download time of the website, or privacy concerns

III. MODEL CHECKING TO ANALYSE THE EVENT LOGS AND E-COMMERCE WEBSITES

Linear Temporal Logic and Model Checking: We are considering a program state in terms of Boolean formulas over a set of atomic propositions. Temporal logics have been used for evolutions. 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 [5]. Having the finite set of possible program executions allows the analysis of the program behaviour. Model checking techniques have been developed to carry out. These techniques check the truth of a set of behavioural specifications, stated in terms of temporal logic formulas, which is composed of the set of executions. Linear Temporal Logic, which defines a logic for traces corresponding to program executions [6]. A trace can be considered as the run of a program, where the setoff atomic propositions corresponds to the set of events or event attributes. 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 [8]. 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 [7]. 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. 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 occurs, for instance. Some of these actions correspond to events that are common to any ecommerce website such as the ones related to visiting the sections containing products. 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.

IV. PROPOSED SYSTEM

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. We propose here a methodology for using it in structured ecommerce websites. The goal is to analyse the usage of ecommerce websites and to discover customer's complex behavioural patterns by means of checking temporal logic formulas describing such behaviours against the log model. At the beginning, web server logs are pre-processed to extract the detailed traces. 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 wish list, 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.

V. ARCHITECTURE DIAGRAM

Admin first login to the website and add the categories of products. Admin can add the sub-categories of the product. Admin can upload picture of the product with all the details. Admin can view the user's details and can view all clustering details. Admin can view user page spending time. Admin can view how much time user is spending on one page and in what product user is interested. Admin can also see the products which are wish listed by the user. Admin can view the purchasing details. Admin can view all the list of purchased product by the users.

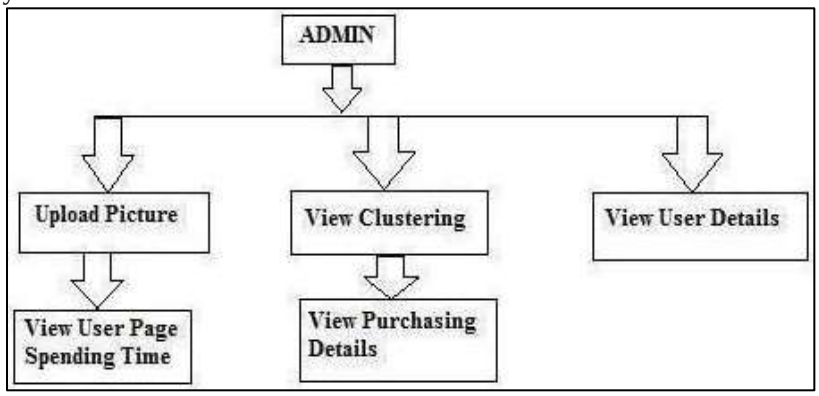

Fig. 1(a):

User first registers their account and login their account. Next search the product and view the product details. User can add the interested product to their wishlist. User can add the interested product to the cart directly from the wishlist anytime. User can buy the product directly from wishlist also. User can also see the other user's feedback about the product. User can see the buying details anytime they want. User can add their feedback about the purchase after they receive the product.

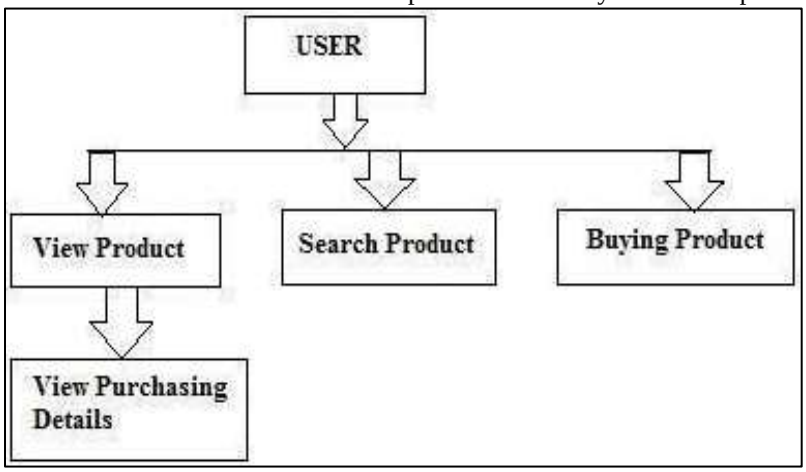

Fig. 2(b):

VI. DATA PREPROCESSING

The initial step of web usage mining analysis is data pre-processing. The raw data have relatively low business value unless they can be transformed and processed to produce actionable knowledge. Therefore, in order to enable the analysis, raw logs must be pre-processed to discard uninteresting requests, to identify user sessions and to prepare the log to enable its analysis. The first two are common to any web usage mining project. 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.
- Deleting requests with erroneous status codes (4xx and 5xx 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. 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).

B. 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 on the other hand, in the simplification sub-phase, log contents are reduced to increase the effectiveness of the model checking techniques.

Other interesting issue is that we cannot identify the category and subcategory to which a product belongs by using the web server logs. 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. 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 are to extract information about user's behaviour 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. 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: First, the user reloads the web page or repeats the same click. Events corresponding to visiting different pages of a give 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.

Second, 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 different pre-processing stages.

VII. CONCLUSION

In the case of open systems, where the sequences of interactions 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 under fitting flower models, from which little interesting information can be extracted. In the paper we apply LTL-based model checking techniques to analyse ecommerce web logs. Plan to extend the set of studied patterns in order to analyse more behavioural patterns and to facilitate their automatic discovery. 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 analyse customer's feedbacks in order to recommend products.

REFERENCES

[1] G. Liu, T. T. Nguyen, G. Zhao, W. Zha, J. Yang, J. Cao, M. Wu, P. Zhao, and W. Chen, "Repeat buyer prediction for e-commerce," in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 155–164.
[2] J. K. Gerrikagoitia, I. Castander, F. Reb´on, and A. Alzua-Sorzabal, "New trends of intelligent e-marketing based on web mining for e-shops," Procedia-Social and Behavioral Sciences, vol. 175, pp. 75–83, 2015.
[3] S.D. Bernhard, C.K. Leung, V.J.Reimer, and J.Westlake, "Clickstream prediction using sequential stream mining techniques with markov chains," in
[4] Proceedings of the 20th International Database Engineering & Applications Symposium, ser. IDEAS '16. New York, NY, USA: ACM, 2016, pp. 24–33. [4] (2017) Google analytics. Accessed 22nd May 2017. [Online]. Available: https://analytics.google.com/analytics/web/
[5] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-streamdata, "Electronic Commerce Research and Applications, vol. 14, no. 1, pp. 1 – 13, 2015.
[6] A. Duret-Lutz, A. Lewkowicz, A. Fauchille, T. Michaud, E. Renault, and L.Xu,"Spot2.0—a frame work for LTL and ω-automata manipulation," in Proceedings of the 14th International Symposium on Automated Technology for Verification and Analysis (ATVA'16), ser. Lecture Notes in Computer Science, vol. 9938. Springer, Oct. 2016, pp. 122–129.
[7] S. Kim, J. Yeo, E. Koh, and 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, ser. WWW '16 Companion. International World Wide Web Conferences Steering Committee, 2016, pp. 57–58.
[8] G. Neelima and S. Rodda, "Predicting user behavior through sessions using the web log mining," in 2016 International Conference on Advances in Human Machine Interaction (HMI), 2016, pp. 1–5.
[9] J. Qi, Z. Zhang, S. Jeon, and Y. Zhou, "Mining customer requirements from online reviews: A product improvement perspective," Information & Management, vol. 53, no. 8, pp. 951 – 963, 2016.
[10] Y. Kang and L. Zhou, "Rube: Rule-based methods for extracting product features from online consumer reviews," Information & Management, vol. 54, no. 2, pp. 166 – 176, 2017.