Business Process Mining from E-Commerce Web Logs

Nicolas Poggi[1,2], Vinod Muthusamy[3], David Carrera[1,2], and Rania Khalaf[3]

[1] Technical University of Catalonia (UPC) Barcelona, Spain
[2] Barcelona Supercomputing Center (BSC) Barcelona, Spain
[3] IBM T. J. Watson Research Center Yorktown, New York, USA

Abstract. The dynamic nature of the Web and its increasing importance as an economic platform create the need of new methods and tools for business efficiency. Current Web analytic tools do not provide the necessary abstracted view of the underlying customer processes and critical paths of site visitor behavior. Such information can offer insights for businesses to react effectively and efficiently. We propose applying Business Process Management (BPM) methodologies to e-commerce Website logs, and present the challenges, results and potential benefits of such an approach.

We use the Business Process Insight (BPI) platform, a collaborative process intelligence toolset that implements the discovery of loosely-coupled processes, and includes novel process mining techniques suitable for the Web. Experiments are performed on custom click-stream logs from a large online travel and booking agency. We first compare Web clicks and BPM events, and then present a methodology to classify and transform URLs into events. We evaluate traditional and custom process mining algorithms to extract business models from real-life Web data. The resulting models present an abstracted view of the relation between pages, exit points, and critical paths taken by customers. Such models show important improvements and aid high-level decision making and optimization of e-commerce sites compared to current state-of-art Web analytics.

1 Introduction

To remain competitive, online retailers need to adapt in an agile, non-structured way, resulting in large, unstructured websites and rapidly changing server resource demands [14]. Moreover, Conversion Rates (CR), the fraction of users that reach a certain goal, such as buying a product on the site, are decreasing: less than 2% of visits result in a purchase on most sites [14]. A low CR is influenced by factors including affiliation programs, changes in user habits such as comparing different sites at the same time [15], and meta-crawling. For example, *Kayak.com* and similar meta-crawlers present the user the best results gathered from several sites, thereby lowering the visits to each site and the CR.

Most online businesses rely on free Web analytic tools to inform their Web marketing campaigns and strategic business decisions. However these tools currently do not provide the necessary abstracted view of the customer's actual

F. Daniel, J. Wang, and B. Weber (Eds.): BPM 2013, LNCS 8094, pp. 65–80, 2013.

behavior on the site. Without the proper tools and abstractions, site owners have a simplified and incorrect understanding of their users' real interaction patterns on the site, and how they evolve.

In this paper we apply Business Process Management (BPM) methodologies to e-commerce Website logs. Structured formal models of user behavior can provide insights on potential improvements to the site. In particular, providing a high-level abstracted view of the workflows leading to purchases and most common exit pages in order to make decisions on site optimization. BPM concerns the management of business processes including the modeling, design, execution, monitoring, and optimization of processes [8]. While loosely-structured to completely ad-hoc processes have not traditionally not been considered by BPM, we (and others [7]) see this is part of a spectrum [19].

Unlike Web analytics [9], process analytics is concerned with correlating events [20], mining for process models [24,26,18], and predicting behavior [25]. We propose treating a user's web clicks as an unstructured process, and use process mining algorithms to discover user behavior. The mined process model captures the causality and paths of user interactions that lead to certain outcomes of interest, such as buying a product. Such insights can be difficult to extract from traditional Web analytic tools.

We use the Business Process Insight (BPI) platform, a collaborative process intelligence toolset [19]. BPI includes the knowledge-based process miner, which differs from traditional process mining in its initial search structure and the set of activities considered for edge operations.

We use a real data set from Atrapalo, an online travel and booking agency (OTA) that includes popular services such as flight and hotel reservation systems. The data set includes the HTTP requests made by customers to the site over a three month period, captured using Real User Monitoring techniques. We apply process analytics to this dataset, and make three main contributions:

1. We outline how to transform web clicks into tasks suitable for analysis and modeling with BPM tools. In particular, we classify the URLs that correspond to web clicks into high level tasks. We compare both a manual classification approach with knowledge from a domain expert, and an automatic classification algorithm. The tasks are then grouped into web sessions representing a particular customer's interaction with the site.
2. We describe how to mine business processes that includes how regular web visitors and customers behave. A challenge here is that, by design, most process mining algorithms capture only the most common behavior in order to keep the resulting mined process model simple enough for a human to understand. However, in web commerce data, the behaviors of interest, such as a customer buying a product, are infrequent. We address this issue with techniques such as saturating the dataset with low frequency behavior we wish to observe, clustering the process instances to extract patterns of behavior, and using a knowledge-based processing mining algorithm.
3. We evaluate the use of the knowledge-based mining algorithm under a variety of conditions, and explain its suitability to extract process models that