On the Concept of Automatic User Behavior Profiling of Websites

Abhimanyu Panwar, University of Alberta, Edmonton, Canada
Iosif-Viorel Onut, IBM Canada, Ottawa, Canada
Michael Smith, University of Calgary, Calgary, Canada
James Miller, University of Alberta, Edmonton, Canada

ABSTRACT

User behavior profiling of websites can provide an operator with an estimate of what is actually transpiring on their site. This type of information is essential to keep ahead of the curve in a commercial environment where competition is extremely fierce and continuously evolving. The authors present an automated methodology that uses economically available web server logs to mine User Behavior Profiles (UBP) without adding significant overhead to an existing web system. They prepare user traces from the log files based on the 35 most common actions found on popular websites, and 9 user behavior profiles which describe the majority of current activity patterns identified from those sites. They classify the user trace into a UBP via a Hidden Markov Model (HMM) based classification approach. The authors applied this methodology to the logs of a virtual e-commerce website, and an industrial case study to demonstrate the validity of the proposed approach.

KEYWORDS

E-Commerce, Experimentation, Hidden Markov Models, Industrial Case Study, Modeling, User Behavior, User Sessions

1. INTRODUCTION

Working under competitive circumstances, websites have to offer an optimum service to sustain user interest. Understanding the browsing behavior of users has the potential to boost visit-to-purchase conversion rates above the current norm of 2% (Dave, 2015). However, the sheer scale of web traffic presents challenges and opportunities when attempting to identify potential areas of improvements in the delivery of services. We propose a low-overhead methodology to automatically extract a list of User Behavior Profiles (UBP). It is argued that these UBPs correlate with the “how, why and for what” of the semantics and services offered by a website.

Tools that support such a methodology can be usefully exploited in many different ways. Website business managers can make informed decisions about refinements in the business processes, practices and services offered. Software developers may focus on and optimize those parts of the code most heavily used during purchase activities. Scanner tools can be designed to better simulate attacker activities based on this pre-determined domain knowledge of the website. Security testing can be

DOI: 10.4018/IJSSOE.2017010101

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
performed by web application security scanners. However, this requires manual intervention by a security expert, at various points in its lifecycle to overcome the scanner’s inability to accurately crawl “deep” (Doupé et al., 2010). Pre-determined domain knowledge from the automatically inferred UBPs allows a security expert to become more familiar with the key functionality of a site, and hence allows them to maximize the effectiveness of their manual intervention.

In many websites, the pre-processing of the user’s requests, e.g. form validation, takes place on the client side. However, all of the important and crucial requests are communicated to and logged by the server. The goal of this study is to present a methodology to (1) infer a list of prominent user behaviors by mining these server logs; and (2) update the profiles as user behaviors evolve and change with time. Both of these tasks must be accomplished while only introducing a minimum additional overhead on the system.

The main contributions of this study are:

1. Traditionally, site profiles are only understood from a static perspective. We present a novel problem of finding user behaviour profiles of a website by exploiting the server logs to provide a dynamic aspect to profiling which promises to provide great benefits;
2. We conduct an empirical study on popular websites to create an alphabet of labels representing a set of common actions related to a set of webpages. This alphabet provides syntax for describing usage patterns in terms of fundamental building blocks;
3. We conduct an empirical study on popular websites to create a list of the most commonly utilized user behaviour profiles. Each profile represents a sequence of webpages requested by the user to fulfil a purpose while browsing the website;
4. We present an automated system which mines the user behaviour profiles of a website by exploiting the server logs. We model the browsing behaviour of users using a Hidden Markov Model (HMM) approach, and experimentally establish that this technique is superior to alternative algorithms;
5. We demonstrate the effectiveness of the proposed methodology by performing case studies on a virtual and a small-scale industrial e-commerce website.

The rest of the study is organized as follows: Section 2 presents various definitions. Section 3 presents the alphabet of labels, list of UBPs, inferred from the study; it also presents the methodology for estimating these UBP from the server logs. We introduce an experimental dataset; and the results from analyzing the dataset are given in Section 4. We present an industrial case study using the proposed methodology in Section 5. A discussion about the viability of the approach is given in Section 6. We present related work in Section 7. Threats to validity of the experiment are provided in Section 8. Section 9 provides some concluding remarks.

2. DEFINITIONS

User Session: A set of requests performed by a user (of a website) in a “continuous” time period. A financially-oriented web site will have a policy to close a user session after a certain time of inactivity.

User Trace: A time ordered sequence of requests, \( T = \left( r_1, r_2, r_3, \ldots, r_p \right) \) within a single user session from a website’s server of the form:

www.eshopping.com
www.eshopping.com/category1/item1
www.eshopping.com/login
www.eshopping.com/user1234/profile
www.eshopping.com/user1234/cart
www.eshopping.com/payment
www.eshopping.com/logout)
Related Content

Backstepping Control for Synchronizing Fractional-Order Chaotic Systems
www.igi-global.com/chapter/backstepping-control-for-synchronizing-fractional-order-chaotic-systems/204799?camid=4v1a

Supporting Consistency during the Development and Evolution of Quality Unified Use-Misuse Case Models
www.igi-global.com/article/supporting-consistency-during-the-development-and-evolution-of-quality-unified-use-misuse-case-models/142038?camid=4v1a
An SMT-based Approach for Generating Coverage Oriented Metamodel Instances
[www.igi-global.com/article/an-smt-based-approach-for-generating-coverage-oriented-metamodel-instances/170518?camid=4v1a](www.igi-global.com/article/an-smt-based-approach-for-generating-coverage-oriented-metamodel-instances/170518?camid=4v1a)

A Novel Approach for Ontology-Based Dimensionality Reduction for Web Text Document Classification
[www.igi-global.com/article/a-novel-approach-for-ontology-based-dimensionality-reduction-for-web-text-document-classification/187171?camid=4v1a](www.igi-global.com/article/a-novel-approach-for-ontology-based-dimensionality-reduction-for-web-text-document-classification/187171?camid=4v1a)